THE UNIVERSITY
of EDINBURGH

A Mathematical Model in Absolute Units for the *Arabidopsis* Circadian Oscillator

José María Uriel Urquiza García

PhD Thesis

Supervisor: Prof. Andrew J. Millar

THE UNIVERSITY *of* EDINBURGH

PhD in Structural and Molecular Biology, The University of Edinburgh, 2017

Declaration of originality

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise. The work has not been submitted for any other degree or professional qualification

José María Uriel Urquiza García. May 2018

Acknowledgements

I chose the University of Edinburgh 6 years ago for pursing a MSc in Systems and Synthetic Biology. I was clueless of the endeavour I was starting. Edinburgh is a tremendous place for doing science and it has become a second home. The scientific endeavour is emotionally challenging, the first person I want to acknowledge is the support from Luis Fernando Montaño through all the time we have spent in Scotland doing our corresponding PhD studies. It would have been a lonelier endeavour without its company during this past five years. We already collected plenty of anecdotes and hopefully many more to come. I want to thank all the members of the Millar lab. They played important roles during my doctoral work. Specially Dr Daniel Seaton for helping me starting up my PhD project. However, every single member contributed somehow to the final outcomes of my PhD with suggestions, discussion and trouble shooting. The members of the staff specially Julie and Eliane from Waddington, Sophie and Pat in Rutherford (No transgenic plants without them). I also include the members of the different labs in Waddington for making it a fantastic research environment. I also want to say thanks to my deceased mother Selina Garcia which has always been a role model of integrity, justice and equity. My father Juan Urquiza shared these values with my mother has been fundamental supporter, cheering up all the way through my academic career. My brother Dr. Juan Humberto Urquiza-García has always been a source of wisdom and important influence on my academic career. Genomics was a perfect fit for me also Scotland and I am thankful for these two bits of information. My sister has been fundamental for developing my emotional resilience crucial when pursing a PhD degree far from home, in a language I thought I understood, she is grate at spotting the ironic elements in many painful processes, this document is finally done older sister. My grandparents have been a second pair of parents both Aurora Chavez and Raúl García Araujo did a tremendous job supporting me in critical moments in my life, including my PhD. Nuri Serrano Montero for being always there for me in moments of deep need. Also, I want to acknowledge my friends in Edinburgh, for being welcoming and understanding always specially Sarah Heath. This Includes the Mexican community and my British friends for helping me finding my way through the left side of the road. Eilidh Smith, you arrived just at the precise moment to make things very smooth. I want to strongly acknowledge the role of my supervisor Professor Andrew J. Millar. The work presented in this thesis is also consequence of remarkable vision and supervision from him. Andrew has been there supporting me in all the crucial forking points during this work. His expertise in the field is present throughout this document. Working along him has been the most challenging and satisfying experience of my life and I am in positive money with him. Finally, to Consejo Nacional de Ciencia y Tecnología CONACYT (Mexico) in combination with the School of Biological Sciences from the University of Edinburgh for funding the PhD project.

Abstract

The Earth's oblique rotation results in changes in light and temperature across the day and time of year. Living organisms evolved rhythmic behaviours to anticipate these changes and execute appropriate responses at particular times. The current paradigm for the biological clocks in several branches of life is an underlying biochemical oscillator mainly composed by a network of repressive transcription factors. The slow decay in their activity is fundamental for generating anticipatory dynamics. Interestingly, these dynamics can be well appreciated when the biological system is left under constant environmental conditions, where oscillation of several physiological readouts persists with a period close to 24 hours, hence the term "circadian clocks", circa=around dian=day.

In plants the model species *Arabidopsis thaliana* has served as an invaluable tool for analysing the genetics, biochemical, developmental, and physiological effects of the oscillator. Many of these experimental results have been integrated in mechanistic and mathematical theories for the circadian oscillator. These models predict the timing of gene expression and protein presence in several genetic backgrounds and photoperiodic conditions.

The aim of this work is the introduction of a correct mass scale for both the RNA transcript and protein variables of the clock models. The new mass scale is first introduced using published RNA data in absolute units, from qRT-PCR. This required reinterpreting several assumptions of an established clock model (P2011), resulting in an updated version named U2017. I evaluate the performance of the U2017 model in using data in absolute mass units, for the first time for this clock system.

Introducing absolute units for the protein variables takes place by generating hypothetical protein data from the existing qRT-PCR data and comparing a data-driven model with western blot data from the literature. I explore the consequences of these predicted protein numbers for the model's dynamics. The process required a meta-analysis of plant parameter values and genomic information, to interpret the biological relevance of the updated protein parameters. The predicted protein amounts justify, for example, the revised treatment of the Evening Complex in the U2017 model, compared to P2011. The difficulties of introducing absolute units for the protein components are discussed and components for experimental quantification are proposed.

Validating the protein predictions required a new methodology for absolute quantification. The methodology is based on translational fusions with a luciferase reporter than has been little used in plants, NanoLUC. Firstly, the characterisation of NanoLUC as a new circadian reporter was explored using

the clock gene BOA. The results show that this new system is a robust, sensitive and automatable approach for addressing quantitative biology questions.

I selected five clock proteins CCA1, LHY, PRR7, TOC1 and LUX for absolute quantification using the new NanoLUC methodology. Functionality of translation fusions with NanoLUC was assessed by complementation experiments. The closest complementing line for each gene was selected to generate protein time series data. Absolute protein quantities were determined by generation of calibration curves using a recombinant NanoLUC standard. The developed methodology allows absolute quantification comparable to the calibrated qRT-PCR data. These experimental results test the predicted protein amounts and represent a technical resource to understand protein dynamics of Arabidopsis' circadian oscillator quantitatively.

The new experimental, meta-analysis and modelling results in absolute units allows future researchers to incorporate further, quantitative biochemical data.

Chapter 1
Introduction

The earth rotates with a period of 24 hours generating rhythmic changes in light and temperature. Its tilt and orbit around the sun produces seasonal changes with differential latitudinal strength, the most obvious being changes in the duration of the day (photoperiod). Living organisms evolving in this periodic environment possibly resulted in the selection of anticipatory rhythmic behaviours. In particular rhythms with periodicity of 24 hours have been documented in a myriad of living forms (Bünning 1964). These rhythms were original documented for the leaf opening of *Tamarindus indica* by Androsthenes of Thasos during the marches of Alexander the Great. The French astronomer Jean Jaques d'Ortous de Mairan observed the continued opening and closing of *Mimosa pudica* leaves in constant darkness, at the correct external time, with a daily cycle (Figure 1.1A). This provided the first documented persistence of rhythmicity in constant conditions, but several centuries had to pass before better control on experimental settings ruled out hidden *zeitgebers* (German for time givers) as the drivers of rhythmic changes. Franz Halberg in 1959 named these subset of biological rhythms, circadian, with Latin roots "circa" (around) and "dies" (day) (Bünning 1964; McClung 2006). Another important feature observed in circadian rhythms is the presence of temperature compensation. It is expected that as the temperature

increases the period of simple chemical oscillators decreases, however circadian rhythms present temperature surprising resistance to temperature changes possibly emerging from temperature compensating mechanisms (Hastings & Sweeney 1957). Experiments using RNA synthesis inhibitors suggested an underlying genetic component for these rhythms. Konopka and Benzer pioneered genetic experiments in *Drosophila melanogaster*, resulting in the isolation of the first group of circadian mutants named (per^{i}, per^{s}, per^{0}) (Konopka & Benzer 1971). Extensive work on these mutants allowed the isolation of the first clock gene and provided the foundations of modern circadian molecular genetics in Drosophila and other model genetic systems of different phylogenetic branches (animals, bacteria, fungi and plants).

A

B

Figure 1.1 Two examples of rhythmic behaviour in plants in day to day life. A) The leaves of Mimosa pudica remain open across the day and close during the night. B) Flower opening in Papaver somniferum presents rhythmic opening with petals opening in the morning and closing in the evening.

The transcriptional and translational feedback loop a basic structure for the generation of circadian rhythms.

Excellent and extensive treaties and reviews exist for the clock mechanisms described in the following section. However, I provide a general overview of circadian networks and an intuitive interpretation of how the rhythms emerge in these networks. Also aiming to provide a more solid ground for

understanding the applicability of the work developed in this thesis on other circadian systems.

The drosophila circadian network.

Mapping the locus containing the *per* mutations allowed the use of state of the art molecular approaches for tracking the expression *PERIOD (PER)* gene at the transcriptional and protein level (Bargiello et al. 1984; Anon 1988).

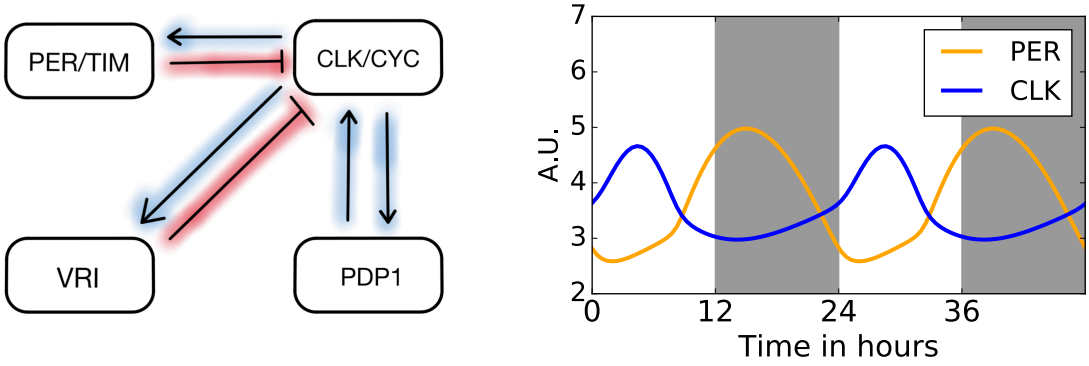


Figure 1.1 Abstraction for the Drosophila circadian network. Red arrows indicate suppression and blue arrows activation. The diagram depicts abstracted components for providing intuition for the underlying network. On the right panel a simulation depicting the dynamics of PER and CLK across diurnal cycles using the model from (Xie & Kulasiri 2007).

These studies showed that the transcript change rhythmically across the day with higher levels of PER protein during night vs day (Figure 1.1). Further work provided evidence for supporting PER protein as a transcription factors (Liu et al. 1992). Integration of this data allowed proposing a model where cycling of PER transcript and protein results in a transcriptional and translational

feedback that with enough delay allows the emergence of oscillatory dynamics. As part of the feedback loop mechanism the proteins CLOCK (CLK) and CYCLE (CYC) form a complex that promote the expression of PER and the TIMLESS (TIM) protein which forms a complex with PER. The PER/TIM complex supresses the activating capacity of CLK/CYC complex impacting on the production of *PER* and *TIM* transcripts. The phosphorylation of TIM by shaggy promotes the transport of the PER/TIM dimmer into the nucleus, therefore it is important in dictating the pace of the oscillator (Martinek et al. 2001).

In addition to the PER/TIM arm in the oscillator, PDP1 act as a transcriptional activator of *CLK* gene providing an extra arm in the network which might contribute to a more robustness of circadian rhythmicity (Cyran et al. 2003). The photoreceptor CRYOTOCHROME (CRY) it is possibly input to the clock. Elimination of this protein results in a lack of proper entrainment to light and dark cycles (Anon 1998). Molecular evidence suggest that it acts as transcriptional repressor (Collins et al. 2006). The placement of this molecular component in the network remains elusive. However, it has been shown that TIM degradation rate is light responsive therefore providing a light input to the network (Zeng et al. 1996). Computational modelling of the network depicted in Figure 1.1, provides theoretical evidence that it is capable of emulating phase of expression and light entrainment properties captured by the phase response curve (PRC) of the system. The phase response curve summarizes

the sensitivity of the circadian system to a particular zeitgeber at different times of circadian cycle. This theoretical results provide a framework that formally tests the network capability of generating circadian rhythms (Xie & Kulasiri 2007).

The mammalian circadian genetic network.

The mammalian clock presents some similarities to the Drosophila system. At the protein level it presents conservation for the PERIOD gene however several gene paralogous are present in the genome. Interestingly, PERs can form heterodimers with CRYPTOCHROMES which is a family of transcription factors. These heterodimers repress the expression of their cognate genes resulting in a negative feedback mechanism (Figure 1.2). Similar to dPER (Drosophila PER) mPERs (Mammalian PER) are present at night (Figure 1.2). In mammals, CLK is also present and forming a dimer with Brain and muscle Arnt-Like protein 1 (BMAL1) which related to the CYC protein from drosophila (Anon 1997). CLK/BML1 dimer acts as an activator across the genome while PERs/CRYs act as repressors (Koike et al. 2012). However, the circadian cycling is provided by interaction of this components with other components including REV-ERBs and RORs (Figure 2.1). Detailed computational modelling of the network has provided as in the Drosophila oscillator evidence that the network can generate oscillations that are compatible with experimental evidence. For example, modelling of the oscillator showed the relevance of stoichiometric balance for robust oscillatory behaviour which

explain the behaviour of relative amplitude in for a plethora of clock mutants (Kim & Forger 2012). Clock models in mammals have used transcript and protein data in relative terms. However, new experimental results for the protein components could allow the introduction of units on the mass scale in future theoretical work (Narumi et al. 2016).

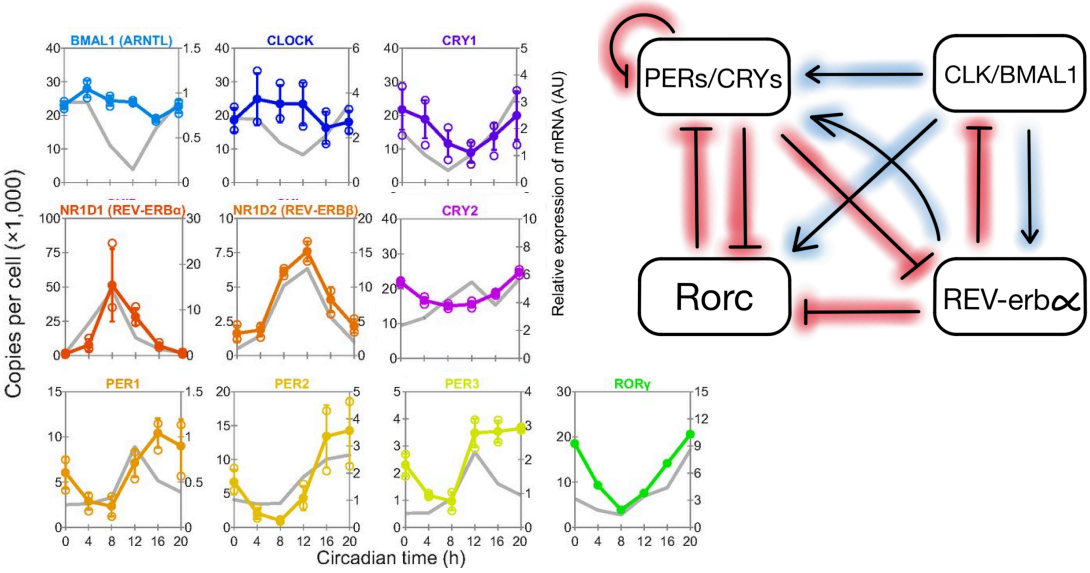


Figure 1.2 Abstraction of the mammalian circadian network with data in absolute units.

Visiting the circadian network of a lower eukaryote *Neurospora crassa*.

Neurospora crassa has been a molecular biology workhorse for linking genotypes to phenotypes. In this system, *frequency(frq)* was the first gene that was shown to fulfil all the properties for being part of a central clock mechanism. *frq* is a morning expressed gene that modulates the activity of

WHITE COLLAR COMPLEX (WCC) by recruiting kinases that hyperphosphorylates the complex resulting in hyperWCC. The latter complex is composed by the protein products of two morning expressed genes, *white collar-1 (wc-1)* and white collar-2 (*wc-2)*. The hyperWCC form is transcriptionally inactive and unable to activate *frq* transcription, closing in this way the feedback loop in the system (Figure 1.3). WCC also binds to the promoter of more than 400 genes of *Neurospora* genome explaining the higher percentage (~20%) of transcripts presenting circadian dynamics (Smith et al. 2010).

.

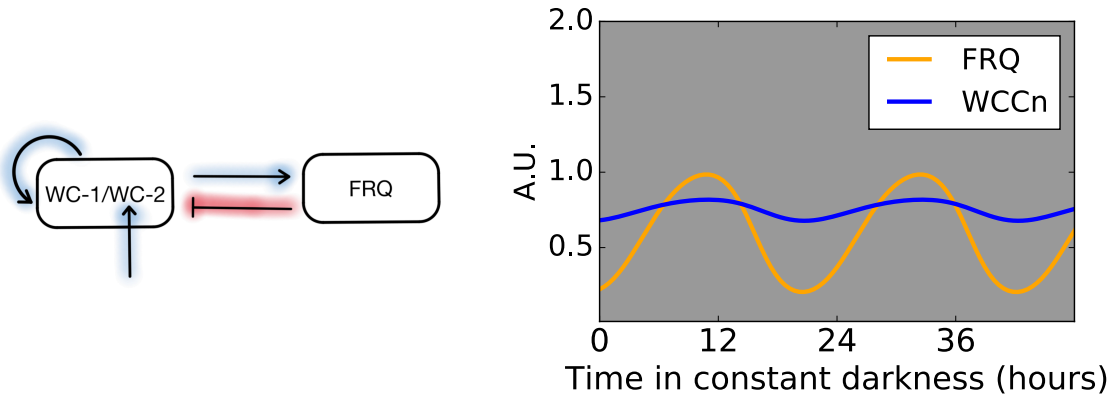


Figure 1.3 The Neurospora network produces rhythmic oscillator in the activity of the WCC (WC-1/WC-2) that impact the expression of hundreds of genes in the genome. WCC activity is suppressed by FRQ mediated kinase recruitment and phosphorylation. Lack of WCC activity results in down regulation of frq transcript and slow accumulation of new non-hyperphosphorylated WCC restarting the cycle, as shown in the left panel (Simulated oscillations using the model of 11 Tseng et al. 2012).

The *Neurospora* system is particularly interesting because temperature impacts the translation of FRQ resulting in two different isoforms as a result of change in translation initiation form different start codons in the *frq* transcript. These are a long and a short version of FRQ, l-FRQ and s-FRQ respectively. Akaman et al built a simple mathematical model incorporating this molecular mechanism and explored the impact of temperature changes on the period of the oscillator. Interestingly, they found that most important control coefficients are linked to the translation rates of l-FRQ and s-FRQ providing theoretical support for the relevance of this switch for temperature compensation in an oscillator (Akman et al. 2008).

In terms of units the models so far developed for this system present arbitrary units, even though FRQ protein was the first clock element to be quantified in absolute units (Merrow et al. 1997). Future work in this direction needs to be performed for incorporating absolute units in mass scale of the models or sequence information, which has been evaluated to some extent for example in the *Bacillus subtilis sin* operon (Voigt et al. 2005).

.

Circadian Rhythms in *Arabidopsis*.

In addition to evident rhythms in leaf and petal movements, other aspects of plant development and physiology have been shown to present circadian

rhythmicity. These include scent emission, stomatal aperture, gated responses to pathogens and pests, hypocotyl elongation and flowering responses. Furthermore, characterization of the temporal dynamics of the expression of this gene and decay of the protein encoded by it provided evidence for a negative feedback loop with delay, as a basic mechanism for the generation of circadian rhythms.

Surprisingly efforts to find clock genes through a homology based approached failed in the identification of possible candidates for clock genes in plants (Millar et al. 1995). Evidence for circadian rhythms in *CHLOROPHYLL A/B-BINDING PROTEIN (CAB)* mRNA in wheat (Nagy et al. 1988) and *Arabidopsis* (Millar & Kay 1991) and state of the art genetic engineering in the early 1990's allowed the generation a synthetic circadian phenotype in *Arabidopsis thaliana*, using the luciferase gene from firefly fused to the promoter of *CAB2* (Millar et al. 1992). This was a ground-breaking technology that literally illuminated the circadian field. In *Arabidopsis*, it allowed the isolation of the first clock mutants that presented miss-expression of the *CAB2:LUC* transgene resulting in a mutant collection named "timing of CAB expression" (*toc*). This collection displayed different circadian phenotypes including the short period mutant *toc1*, which was mapped to the lower arm of chromosome 5, demonstrating the power of the non-invasive *CAB2:LUC* assay (Millar et al. 1995).

However, the first molecular components to be placed as elements in the circadian oscillator were the MYB transcription factors *CIRCADIAN CLOCK ASSOCIATED 1 (CCA1)* and *LATE ELONGATED HYPOCTYL (LHY). CCA1* was identified as a binding activity to a cytosine and adenine rich region CA-1 in the promoter of *CAB1* (Sun et al. 1993; Wang et al. 1997). Expression of *CCA1* provided evidence for its role as an element in the circadian oscillator, as circadian rhythms were affected and also the presence of a negative auto-regulatory loop was observed when measuring the levels of the endogenous *CCA1* transcript {Wang:1998zo}. *LHY* was reported at the same time, where similar evidence placed it as an element in the oscillator. Disruption of circadian rhythms and hints of autoregulation could be observed. But opposite to the down-regulation observed for the native *CCA1* in *CCA1*-ox lines, intermediate levels of WT *LHY* were present in the *lhy* overexpressing mutants {Schaffer:1998wx}.

Subsequent cloning of *TOC1* and analysis closed a loop between *LHY/CCA1* and *TOC1*, which was not possible with CCA1 alone as the profile of this protein could not explain the 24 rhythm (Alabadí et al. 2001). *LHY/CCA1* are morning expressed genes with a limited protein expression profile that followed the RNA dynamics. Neither CCA1 transcript of protein were present in the early night. *TOC1* cloning and characterisation of its transcript by Northern blot analysis showed that it is an evening expressed gene and in antiphase to *LHY* and *CCA1* transcripts (Strayer et al. 2000). The elucidation of *TOC1* as a

response regulator provided motivation for exploring the temporal expression of other members of a gene family denominated PSEUDO RESPONSE REGULATORS (PRRs). This family of genes were independently identified when studying Histidine Kinase Response Regulators. However, the PRR members were classified together because they lack the canonical key histidine residue involved in the phosphorelay mechanisms that Response Regulators present in plants and bacteria (Matsushika et al. 2000).

At this stage *TOC1* was originally proposed as an activator of *LHY* and *CCA1* and the latter two genes working as repressors of *TOC1.* CCA1-OX lines resulting in down regulation of TOC1 transcript providing evidence of CCA1 repressive activity. TOC1 was originally classified as an activator because *toc1-2* a hypomorphic recessive allele results in down regulation *of CCA1* and *LHY* transcripts*. toc1-1* an antimorphic allele does not present changes in levels of CCA1 and LHY even though the period is shorter compared to wild type (Alabadí et al. 2001). This closed the first loop for the circadian oscillator where TOC1 could bring enough delay for generating circadian rhythms in plants. Interestingly these two genes are also preserved in the smallest free-living organisms, the algae *Ostreococcous tauri* (Noordally & Millar 2015). Suggesting an ancient mechanism that is present in the green lineage. In the era of molecular genetics, a broad range of experimental evidence showed that in order to understand circadian rhythmicity dynamical models are need

for a gaining understanding the role elements in the circadian mechanisms play for generating the observed oscillatory behaviour.

Paring the evolution of the experimentally determined clock network and its formal, mathematical representation.

The LHY/CCA1-TOC1 loop was formalised originally with a simple mathematical model, using a set of ordinary differential equations that represents RNA and protein of LHY/CCA1 and TOC1 (Figure 1.2A). LHY and CCA1 are collapsed in a single RNA or protein variable (cL_m, cL respectively) as these genes have been shown to be partially redundant, while TOC1 had variables cT_m, cT, where c stands for concentration and m denotes the transcript variable. In this model, a hypothetical protein *cP* was first introduced as a biological light regulator in the clock, allowing light entrainment in the oscillator (Locke et al. 2005). The ideal dynamics of this variable is a component that accumulates during the night and then promotes acute light activation at the transition from darkness to light. The acute activation is provided by cP only being active in the presence of light and presenting very strong degradation only in the light regime. At this moment linking the clock model to RNA time series was very challenging as collecting RNA in plants in a time-consuming process that is subjected to strong experimental noise given the nature of the plant tissue. Several steps are need in order to obtain good quality RNA molecules. In order to obtain parameters for the model *ad hoc*

fitting functions were developed using the noisy RNA time series. Resulting in the first model for the plant circadian clock.

It is clear in the topology of the LHY/CCA1-TOC1 network that the *lhy/cca1* double mutant cannot not sustain oscillations (Figure 1.2A). This double mutant is important because the *toc1-2* present short period in constant light. This could be explained by the presence of a single loop made by CCA1 or LHY. It was already known that overexpression of either of them results in downregulation of the endogenous transcript. Therefore, in order to explain the remaining oscillations a new hypothetical component *Y* was introduced to explain persistent rhythmicity in *lhy/cca1* mutant background (Figure 1.2B). The resulting model constitutes a set of interlocked transcriptional and translational feedback loops. This genetic network architecture resembles to some extent the presence of this interlocked loops from other organisms for e.g. *Drosophila.*

At this point several genes were identified as possible members of the circadian oscillator. A family of *PSEUDO RESPONSE REGULATORS (PRRs)* were shown to present circadian rhythmicity with order of expression PRR9, PRR7, PRR5, PRR5, TOC1(PRR1) (Makino et al. 2000). Overexpression of PRR1(TOC1) results in repression of the PRR family members, fulfilling one requirement for linking them as possible clock members (Makino et al. 2000).

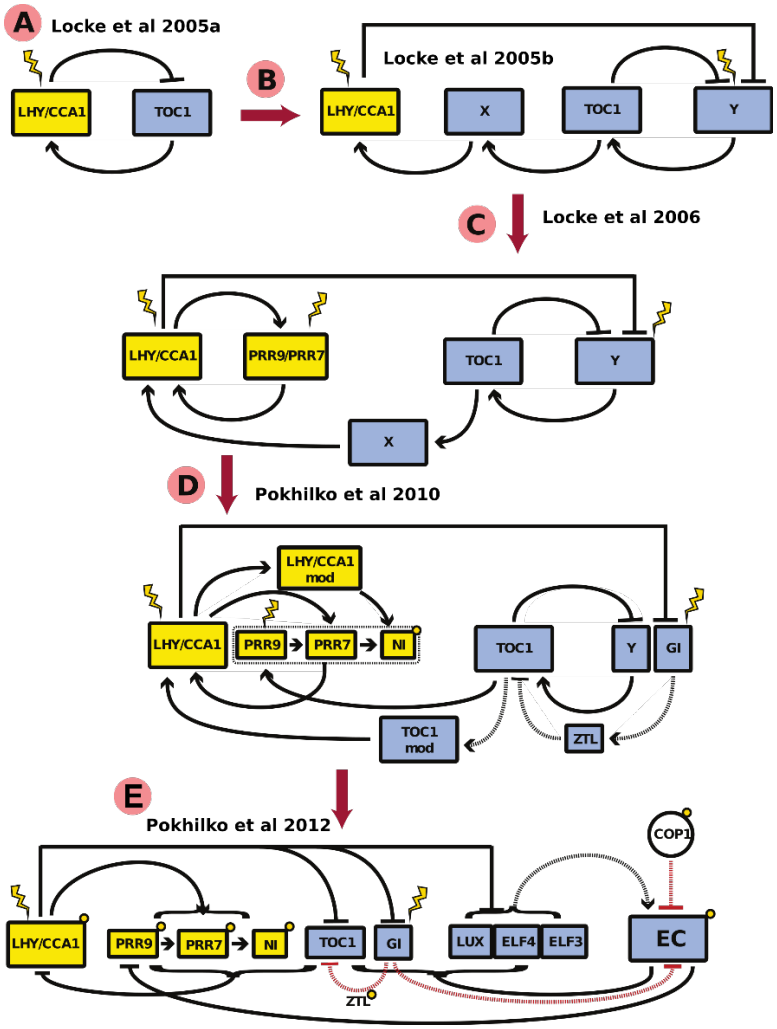


Figure 1.4 Evolution of the mathematical representation of Arabidopsis circadian oscillator. Yellow boxes, morning expressed genes, blue boxes evening expressed genes. Bolts represent light input at transcription. Yellow dots light inputs at the posttranscriptional level.

The striking expression of the PRRs provided a path for a more comprehensive explanation for the sharp expression profile of *LHY* and *CCA1*. In the case of the PRRs the lack of the canonical histidine of RR suggests a different mechanism of action. Nonetheless the PRRs can still be phosphorylated which affects its nucleo-cytoplasmic partitioning (Fujiwara et al. 2008). The PRRs are an interesting group of protein because they link the circadian oscillator with

environmental outputs related to flowering time. They are stabilised in the presence of blue light. They are targeted for degradation by interacting with ZEITELUPE, a LOV/PAS F-box protein that was ben shown to affect *CAB2:LUC* rhythms and biochemical evidence placed it as a regulator of TOC1 protein stability (Somers et al. 2000; Mas et al. 2003). Interestingly, ZTL contains a PAS domain which has been described for *Drosophila* PER protein. The LOV domain of ZTL is responsive to blue light therefore providing an input for the clock by modifying the degradation rate of TOC1 and PRR5 proteins between light and dark. Nonetheless up to this moment no need for introducing this extra PRRs was required. Neither the introduction of other clock related gene e.g. *EARLY FLWERING 3 (ELF3)*. The latter presenting very strong phenotypes for hypocotyl growth and flowering time, which are associated phenotypes in clock mutants (Hicks et al. 1996; Dowson-Day & Millar 1999). Further research provided evidence of aberrant circadian dynamics in these mutants. These last two genes were implicated in the gated responsiveness of by the circadian oscillator (McWatters et al. 2000; Doyle et al. 2002).

The path taken was different given that the mathematical model L2005b suggested a particular pattern for the dynamics of the *Y* component. Time-series of transcript abundance for clock affecting genes using WT and *lhy/cca1* double mutant plants *suggested GIGANTEA (GI)* as a putative gene for the *Y* component (Locke et al. 2006). Further experimental work with a *gi* and *lhy/cca1/gi* triple mutant provided stronger support for *Y/GI* hypothesis, as the

dynamics of the model emulated behaviour of *CAB2:LUC* in the mutant backgrounds, providing supporting evidence for the existence of the feedback loop that explained the residual oscillations in the *lhy/cca1* background. In this model also, a new set of genes *PRR7/PRR9* was introduced as period phenotype of the double mutant prr7/9 presented period lengthening of 6 hours in constant light. This was opposite to the 22 hours period *toc1-2 has.* In this model PRR7/9 were bundled together as a small period increase was observed in the single mutants compared to the strong phenotype of the *prr7,prr9* double mutant (Figure 1.2C) (Nakamichi et al. 2005) (Locke et al. 2006; Zeilinger et al. 2006). This predictive property in the modelling approach supported the power of mathematical integration and analysis for understanding qualitative and quantitative aspects of the circadian oscillator. The model was abstracted as three, interlocking feedback loops (morning, central, evening). However, the effects of *GI* in mutant plants did not perfectly match the predicted functions of *Y*, as noted in the original paper (Locke et al. 2005) and others (Martin-Tryon et al. 2007), so the identity of *Y* was not fully clear.

Extensive literature integration and new experimental evidence resulted in an increase in complexity of models (Pokhilko et al. 2010). In P2010 model, post-translational regulation started to take its place with the explicit introduction of ZTL as a regulator of TOC1 protein, and stabilisation of ZTL by GI. The

introduction of ZTL and GI in the model is a crucial step because it provides a way for modulating the clock waveform in a photoperiodic dependent manner.

The striking wave of PRRs expression across the day required the introduction of a temporally modulated CCA1/LHY protein (Figure 1.2D). A modified (putatively, phosphorylated) form of CCA1/LHY (named *Lmod)* was introduced speculatively, as an effective means to reproduce the *PRR* expression patterns, in the absence of relevant data. CCA1 had been experimentally identified as a phosphoprotein, with CASAIN KINASE 2 (CK2) as its kinase, providing evidence for such mechanism (Daniel et al. 2004). The introduction of the full set of *PRR*s resulted in a model that allowed photoperiodic responses for CCA1/LHY variables, similar to what has been observed in experimentally determined for the levels of *CCA1* transcript. Something relevant about kinases is their conservation across kingdoms and which have been recruited for influencing circadian rhythms. These include CK2, WNK1 and GSK3 (Hindle et al. 2014).

Further experimental work enriched understanding of the posttranslational regulatory layer, resulting in the P2011 model. Several new activities were formally incorporated, in the evening of the circadian cycle (Figure 1.2E). *ELF3, ELF4* and *LUX ARRHYTHMO (LUX)* were incorporated as a crucial loop in the clock, replacing *Y*. These proteins form an Evening Complex (EC) (Nusinow et al. 2011). The exact steps in the complex formation were unknown

and had to be assumed in the model, where ELF3:ELF4 dimer is originally formed, finally interacting with LUX to give the EC. New arrows of posttranslational regulation were also included, for example interaction between GI and ELF3 proteins, and light-sensitive degradation of the EC complexes by CONSTITUTIVE PHOTOMORPHOGENESIS 1 (COP1) (Pokhilko et al. 2012). Interestingly a new abstraction for the clock model took place at this point, with a repressilator as the framework (Figure 1.3).

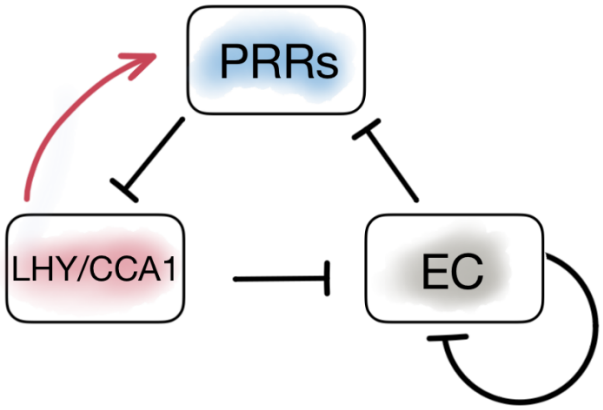


Figure 1.5 First abstraction of the clock mathematical model clarifies the existence of a repressilator (Pokhilko et al. 2012). Arrows represent activation and blunt arrows repression.

The clock in Arabidopsis can be abstracted in the following way. LHY/CCA1 are morning expressed transcription factors which repress the expression of Evening Genes including PRR9, PRR7, PRR5, TOC1, LUX, ELF3 and ELF4, GI. In the morning REVEVILLEs (RVEs) genes that present some degree of conservation with LHY however they activate the expression of *PRR9*, *PRR5* and *TOC1*. The balance of LHY/CCA1 and RVEs determines the wave of expression of the PRRs which in turn feedback to LHY/CCA1 and RVEs,

constraining their activity during the morning. The RVEs form complexes with LINK proteins LNK1, 2, 3 and 4, which is important for RVEs activating properties. The PRRs also regulate each other in a repressive way. TOC1 regulates its own promoter however it also targets PRR9, PRR7 and PRR5. PRR5 acts on PRR7 and PRR7 setting their expression during the morning and the middle of the day. The PRRs are regulated post-translationally by light through targeted degradation of ZTL. The Evening Complex genes ELF3, LUX and ELF4 repress the expression of the PRRs therefore acting across the night and promoting the expression of LHY and CCA1 in an indirect way by down regulating the PRRs.

Realism and abstraction in the clock models

Clock models have focused in producing proper timing prediction in the LD cycle and detailed waveforms for clock component in multiple conditions, light cycles, photoperiods, phase response curves PRC and mutants. However, only relative levels of these components have been used. Other clocks present different levels of characterisation for example the cyanobacterial clock made by KaiA, KaiB and KaiC proteins can be reconstituted in-vitro, offering easier access to its kinetic parameters (Tseng et al. 2014; Lin et al. 2014). However not data or models exist for data in different photoperiods. In the case of the fungal clock there are early absolute quantifications of FREQUENCY (FRQ) protein, from which interesting conclusion were obtained providing evidence for the critical role of FRQ for explaining most of the circadian rhythmicity (Merrow et al. 1997). In the case of the mammalian oscillator currently there

are absolute levels of nuclear concentration of clock proteins (Narumi et al. 2016). Even though a myriad of time series Chip-seq data exist for mice, but no output pathways from the clock have been modelled (Takahashi et al. 2015). Working in absolute units in Arabidopsis offers the opportunity to link a rich source of biological data and clock models for developing the first step towards mathematical modelling of circadian outputs at the genome level.

The emergence of "new" models for the plant oscillator mechanisms and P201x architectures as a paradigm in *Arabidopsis* chronobiology

New clock models have emerged with different objectives in mind. On one hand, efforts have been taken to obtain better fits to experimental data by increasing the wiring complexity and number of components in the F2014 model (Fogelmark & Troein 2014). This relatively new model has a modified interpretation of gene activation in the clock (Figure 1.6), substituting for *Lmod* transcriptional activation by introducing a new network node that represents the *REVEILLE* (RVE) family of transcription factors, in particular *RVE8*. The *RVEs* present similar transcript profile and sequence homology to *LHY/CCA1*. Experimental evidence suggests that they bind to the promoters of *PRR9, PRR5, TOC1* and the evening genes *ELF4* and *LUX*, promoting their expression (Rawat et al. 2011). F2014 also introduced a new gene in the evening, BROTHER OF LUX ARRHYTHMO (BOA, ≡ NOX) as a new element in the evening complex. The authors made a fundamental modification in the architecture. The PRR9 and PRR7 proteins no longer present activating

properties. In P2011, PRR9 -> *PRR7* and in turn PRR7-> *PRR5* (arrow representing activation, in a forward temporal sequence*)*. In F2014, TOC1 --l [*PRR9,7,5*], PRR5 --l [*PRR9,7*], reflecting new data on the targets of PRR proteins. This feature had been previously introduced in the P2013 model (Pokhilko et al. 2013). However, that model was developed to explore the coupling of clock function to the ABA hormone signalling pathway and this intimate connection is less relevant for my work. In F2014, therefore, the phase of expression of the morning *PRRs* are determined by a balance between RVE8 activation during the day and afternoon and evening repression. An interesting feature of this wiring change is the generation of sharper, cuspidate expression profiles for the *PRR* transcripts. This dynamic profile in the *PRRs* was formally explored by (Foo et al. 2016).

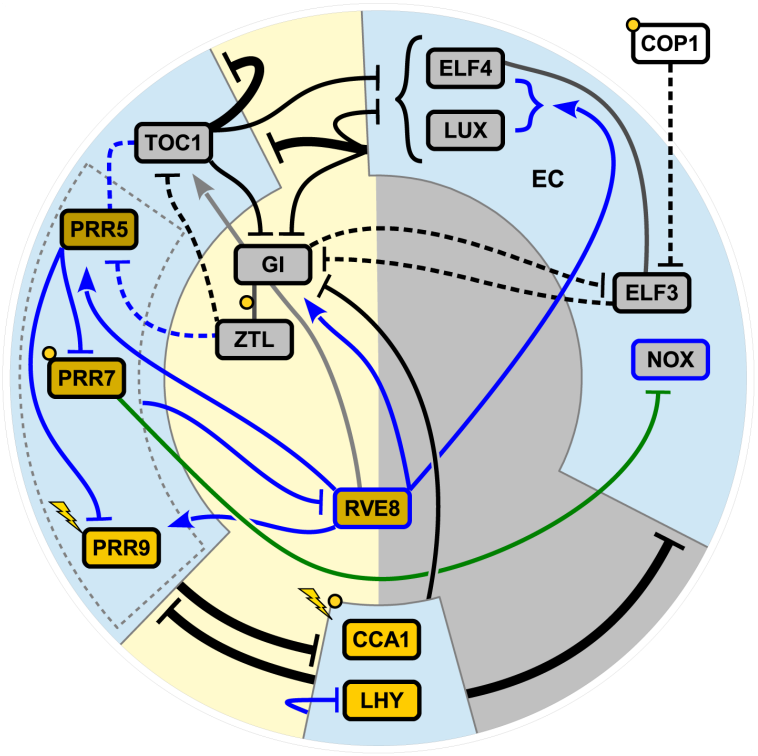


Figure 1.6 Diagrammatic representation of F2014. A substantial increase in model complexity originates from the explicit representation of previously bundled activities like CCA1 and LHY, including dimerization, from new elements like RVE8 and BOA(NOX), and from new connections such as autoregulation of PRRs. F2014 also introduces more complexity at the posttranslational levels with the phosphorylation of TOC1. Reproduced from (Fogelmark & Troein 2014).

F2014 has a substantial increase in network architecture compared to P2011. The authors claimed a better model than P2011, though no model selection criteria were used for this propose. The model limited the number of parameters, by setting all translation rates to 1 and all Hill coefficients to 2. Model selection techniques have not been used for testing current models. The only documented case in the plant clock modelling was an application to test the evening regulation of cold signalling (Keily et al. 2013).

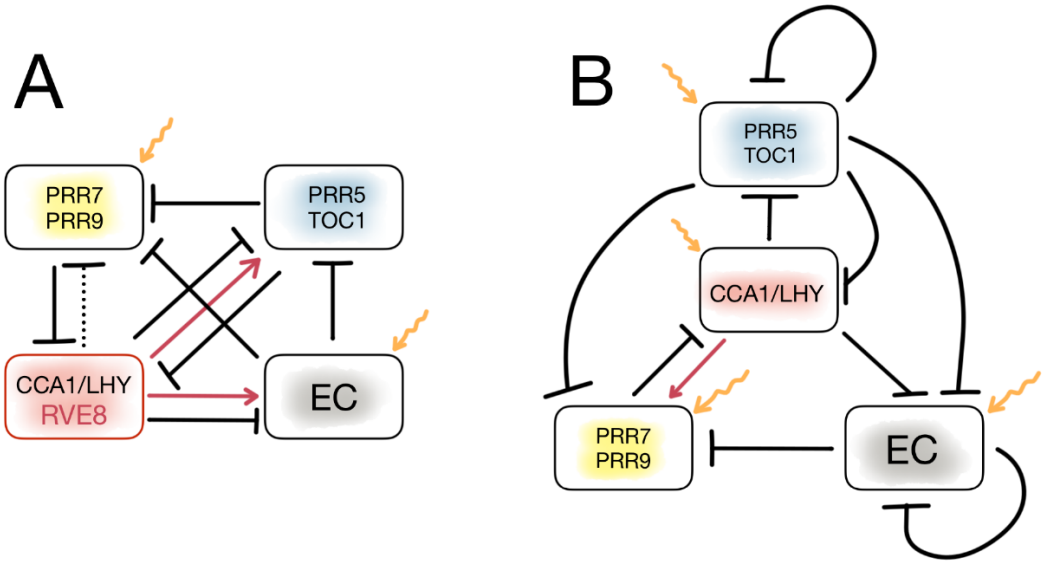


Figure 1.7 Compact representations of the Arabidopsis circadian oscillator. A) The quadripressilator introduced by (Millar 2015) Red arrows represent activation, black blunt arrows repression. Orange arrows represent light inputs. B) Representation proposed by (De Caluwé et al. 2016), similar annotation as for A.

An alternative path is the decrease in model complexity by abstracting the network topology currently available for the clock. This is not a new approach in *Arabidopsis*, as L2005, L2006, P2010 and P2011 are hybrids of explicit and abstracted nodes. Currently two conceptual abstractions exist (Figure 1.7 A,B). These abstractions present similar nodes but some differences in the connections can be appreciated. The Caluwé architecture (B) was explored computationally, showing that it can reproduce to some extent already established qualitative features in the dynamics (De Caluwé et al. 2016).

Reviewing the evolution of clock models shows that the P2011 architecture is a fundamental starting point for future developments. It captures many important circadian features of Arabidopsis clock components. It has also served as a knowledge organising tool, which allows rapid exploration of possible implications of rhythmic control, by users with little experience of the nuts and bolts of the clock mechanisms.

Circadian outputs and modelling efforts in *Arabidopsis*.

The clock controls a myriad of processes. More than 30% of the Arabidopsis genome is under circadian regulation, uncovered by the use of transcriptomic methodologies (Harmer et al. 2000; Edwards et al. 2006; Covington et al. 2008). This supports the idea that the clock is coordinating a myriad of physiological and developmental process. These includes photosynthesis, light and hormone signalling, flowering time, hypocotyl elongation, biotic and abiotic stress (Sanchez & Kay 2016).

Chip-seq experiments have been insightful for showing the pervasive regulation that the clock might exert on transcriptome. It is possible that a large proportion of the circadian oscillation in the transcriptome are a consequence of TF regulation, rather than modulation of RNA stability (Figure 1.8). However, I do not rule out the possibility of this mode of regulation given that no circadian experiments have been performed that interrogate the microRNA component of transcriptome.

Clock models have been used as a framework for further understanding of specific clock inputs or outputs. Salazar et al. described how clock regulation in combination with light signals control the expression of *FKF1* and *CO*. The authors provided alternative models for explaining the transcript dynamics of *FT* transcript (Figure 1.9). This gene encodes key protein that fulfils the role of florigen which promotes the transition from vegetative to reproductive growth in *Arabidopsis* (Kardailsky et al. 1999)

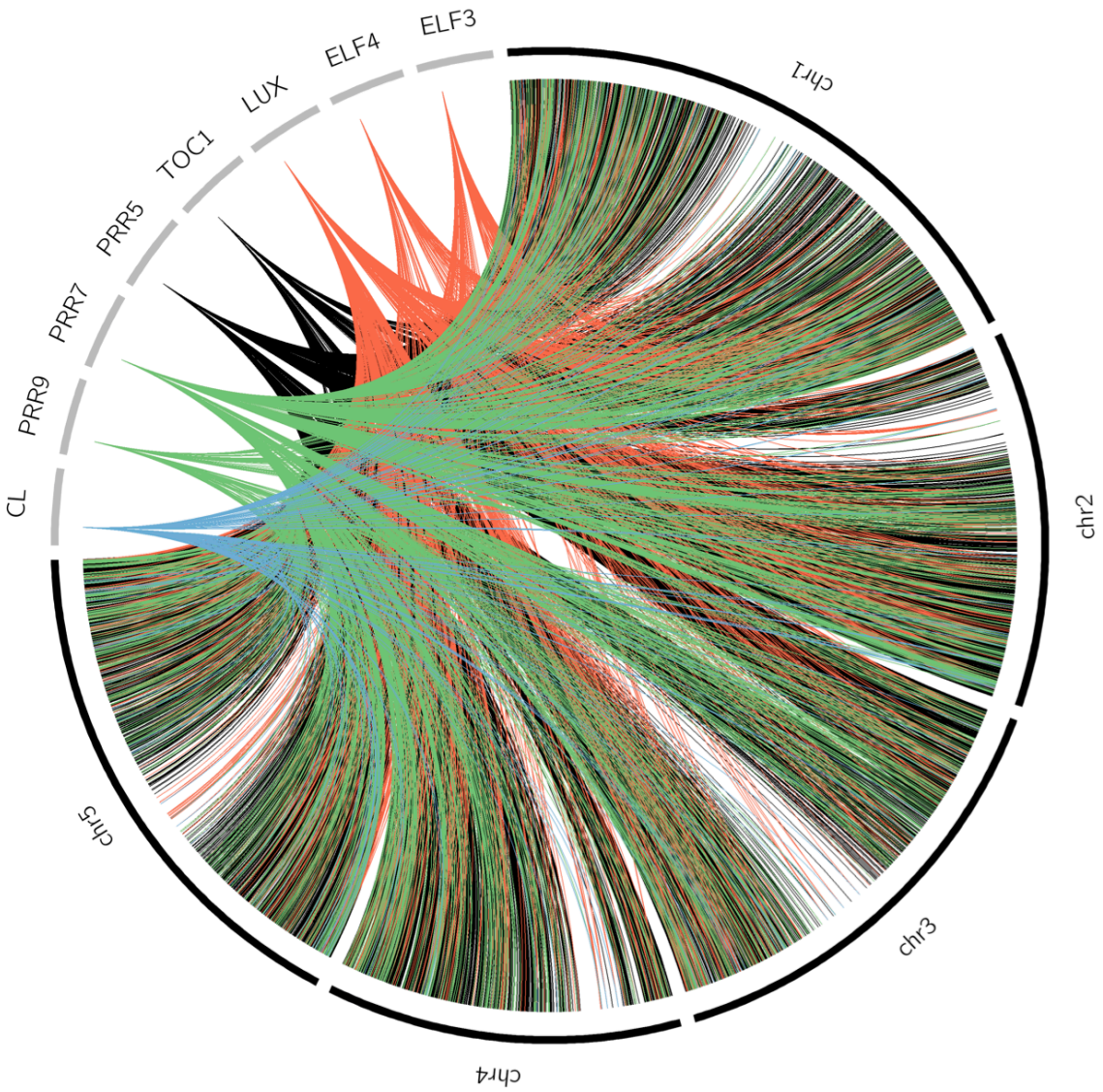


Figure 1.8 Pervasive binding of clock transcription factors in the Arabidopsis Genome. The image was generated in Circos by mapping the ChIP-seq peaks for several clock genes. Regions of low density correspond to centromere regions.

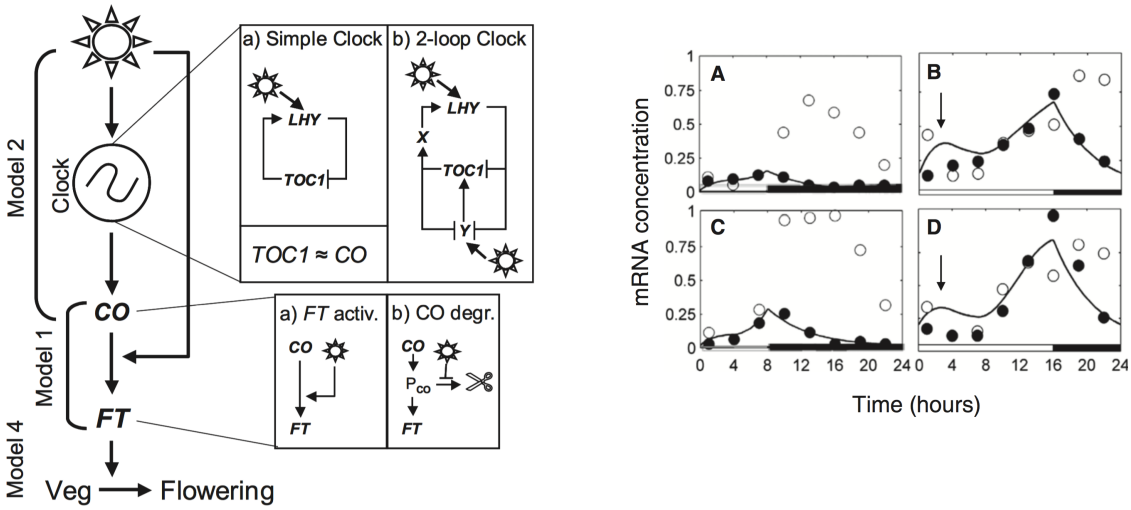


Figure 1.9 Alternative models described for the coincidence mechanism for the regulation of FLOWERING LOCUS T (FT). Two alternative nested models for the clock a and b in combination with two alternative way of regulation of CO protein. Reproduced from (Salazar et al. 2009). Using Model 1a it is possible to reproduce the dynamics of CO (open circles) and FT (closed circles), FT simulation (solid line) in short and long days, upper and lower panel respectively.

This work with FT allowed prediction of light regulated proteins which provide timing information into the photoperiodic pathway of flowering (Song et al. 2012).

Another well documented process that is under the control of the circadian oscillator is growth. Hypocotyl elongation happens at faster rates towards the end of the night which is disrupted in circadian mutants (Nozue et al. 2007). Plenty of experimental work showed that growth is promoter by PHYTOCHROME INTERACAITNG FACTORS (PIFs). They are regulated transcriptionally by the Evening Complex (Nusinow et al. 2011). Light promotes degradation of this protein in PHYTOCHROME B dependent way.

The protein only accumulates at the end of the night, correlating with maximal growth at this point. Therefore, perturbation of clock elements results in growth rate changes. Several clock mutants have hypocotyls with different lengths. For example evening complex mutants present very long hypocotyls (Nusinow et al. 2011).

Extending work in this direction for modelling clock outputs Seaton et al created a model that recapitulates the regulatory for PIFs and several downstream genes including CDF1, IAA29 and ATHB2 which are growth promoting genes. Then, they tested these extensions of the Salazar approach for gaining better understanding of how temperature and photoperiod are interacting with the clock for regulating hypocotyl length and flowering time (Seaton et al. 2015). This work shows that it is possible to link in a gene activity to high order phenotype such as flowering times and hypocotyl elongation in different environmental settings (Figure 1.10).

The last example is the integration of this gene network model into the framework model. This model is the result of incorporating several models including photosynthesis, vernalisation, plant architecture and models of carbon partitioning in Arabidopsis, the aim of this modelling effort allow quantitative predictions of biomass. It incorporates a simple clock model. Therefore, it allows exploring the organismic consequences of perturbing the oscillator (Chew et al. 2017).

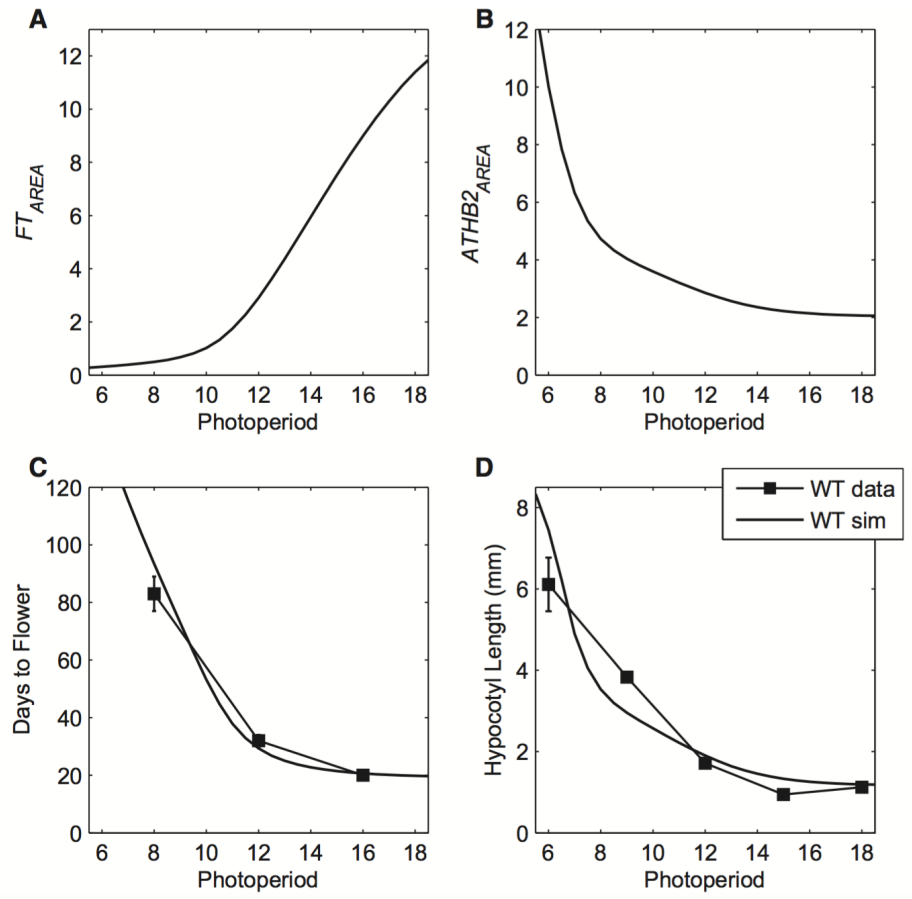


Figure 1.10 Predicting growth and flowering under different environmental conditions. Using the Area under the curve for important regulators of flowering time and hypocotyl elongation the model recapitulates the under different day lengths. The areas for the marker genes were predicted using a modular approach that incorporates the circadian clock and light regulated process. Figure reproduced form (Seaton et al. 2015).

The pervasive phenotypic presence of circadian rhythms suggested a wide regulation of the genome. Important work has been performed to determine which sites in the genome are bound by clock transcription factors by performing Chip-seq experiments. These showed that indeed thousands of genes are bound by clock transcription factors. The provided evidence for the heavy role transcriptional regulation might play in the generation of circadian

oscillation in the transcriptome. However, this Chip-seq experiments are currently snap shots of a dynamic process. In modelling could eventually help to predict the dynamic patterns of binding in different positions and hence transcriptional regulation at these sites. However, this depends in the use of absolute units for unifying experimental observations with theoretical work. Absolute units will provide the tools for unifying sequence information with the dynamics of the system.

Aims of the thesis

-Introducing absolute units for the mass of the state variables and evaluate a P2011 model under this new scenario.

-Obtaining empirical distributions for evaluating the biological soundness of the parameters in absolute units.

-Develop a methodology for performing absolute quantification of plant clock proteins.

-Generate predictions for the expected amount of proteins the plant produces for interpreting the experimental protein quantifications.

-Generation of new transgenic lines for key clock proteins in the mathematical model for performing absolute quantifications.

-Reflect on the new findings and how absolute units help improve modelling circadian rhythms.

The aim of the thesis is evaluating the use of absolute units for finding problems in clock mathematical models.

References

Akman, O. et al., 2008. Isoform switching facilitates period control in the Neurospora crassa circadian clock. *Molecular systems biology*, 4, p.164.

Alabadí, D. et al., 2001. Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science (New York, N.Y.)*, 293(5531), pp.880–883.

Anon, 1988. Antibodies to the period gene product of drosophila reveal diverse tissue distribution and rhythmic changes in the visual system. *Neuron*, 1(2), pp.141–150.

Anon, 1997. cDNA Cloning and Tissue-Specific Expression of a Novel Basic Helix–Loop–Helix/PAS Protein (BMAL1) and Identification of Alternatively Spliced Variants with Alternative Translation Initiation Site Usage. *Biochemical and biophysical research communications*, 233(1), pp.258–264.

Anon, 1998. The cryb Mutation Identifies Cryptochrome as a Circadian Photoreceptor in Drosophila. *Cell*, 95(5), pp.681–692.

Bargiello, T.A., Jackson, F.R. & Young, M.W., 1984. Restoration of circadian behavioural rhythms by gene transfer in Drosophila. *Nature*, 312(5996), pp.752–754.

Bünning, E., 1964. *The Physiological Clock*,

Chew, Y.H. et al., 2017. Linking circadian time to growth rate quantitatively via carbon metabolism. *bioRxiv*, p.105437.

Collins, B. et al., 2006. Drosophila CRYPTOCHROME is a circadian transcriptional repressor. *Current Biology*, 16(5), pp.441–449.

Covington, M. et al., 2008. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome biology*, 9(8).

Cyran, S.A. et al., 2003. vrille, Pdp1, and dClock form a second feedback loop in the Drosophila circadian clock. *Cell*, 112(3), pp.329–341.

Daniel, X., Sugano, S. & Tobin, E., 2004. CK2 phosphorylation of CCA1 is necessary for its circadian oscillator function in Arabidopsis. *Proceedings*

of the National Academy of Sciences of the United States of America, 101(9), pp.3292–3297.

De Caluwé, J. et al., 2016. A Compact Model for the Complex Plant Circadian Clock. *Frontiers in plant science*, 7, p.74.

Dowson-Day, M. & Millar, A., 1999. Circadian dysfunction causes aberrant hypocotyl elongation patterns in Arabidopsis. *The Plant journal : for cell and molecular biology*, 17(1), pp.63–71.

Doyle, M. et al., 2002. The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. *Nature*, 419(6902), pp.74–77.

Edwards, K. et al., 2006. FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. *The Plant cell*, 18(3), pp.639–650.

Fogelmark, K. & Troein, C., 2014. Rethinking transcriptional activation in the Arabidopsis circadian clock. *PLoS computational biology*, 10(7), p.e1003705.

Foo, M., Somers, D.E. & Kim, P.-J., 2016. Kernel Architecture of the Genetic Circuitry of the Arabidopsis Circadian System. L. J. Jensen, ed. *PLoS computational biology*, 12(2), p.e1004748.

Fujiwara, S. et al., 2008. Post-translational regulation of the Arabidopsis circadian clock through selective proteolysis and phosphorylation of pseudo-response regulator proteins. *The Journal of biological chemistry*, 283(34), pp.23073–23083.

Harmer, S. et al., 2000. Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science (New York, N.Y.)*, 290(5499), pp.2110–2113.

Hastings, J. & Sweeney, B., 1957. ON THE MECHANISM OF TEMPERATURE INDEPENDENCE IN A BIOLOGICAL CLOCK. *Proceedings of the National Academy of Sciences of the United States of America*, 43(9), pp.804–811.

Hicks, K. et al., 1996. Conditional circadian dysfunction of the Arabidopsis early-flowering 3 mutant. *Science (New York, N.Y.)*, 274(5288), pp.790–792.

Hindle, M.M. et al., 2014. The reduced kinome of Ostreococcus tauri: core eukaryotic signalling components in a tractable model species. *BMC genomics*, 15(1), p.640.

Kardailsky, I. et al., 1999. Activation tagging of the floral inducer FT. *Science*, 286(5446), pp.1962–1965.

Keily, J. et al., 2013. Model selection reveals control of cold signalling by evening-phased components of the plant circadian clock. *The Plant Journal*, 76(2), pp.247–257.

Kim, J. & Forger, D., 2012. A mechanism for robust circadian timekeeping via stoichiometric balance. *Molecular systems biology*, 8, p.630.

Koike, N. et al., 2012. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science (New York, N.Y.)*, 338(6105), pp.349–354.

Konopka, R.J. & Benzer, S., 1971. Clock mutants of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 68(9), pp.2112–2116.

Lin, J. et al., 2014. Mixtures of opposing phosphorylations within hexamers precisely time feedback in the cyanobacterial circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 111(37), pp.E3937–45.

Liu, X. et al., 1992. The period gene encodes a predominantly nuclear protein in adult Drosophila. *Journal of Neuroscience*, 12(7), pp.2735–2744.

Locke, J. et al., 2006. Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Molecular systems biology*, 2, p.59.

Locke, J., Millar, A. & Turner, M., 2005. Modelling genetic networks with noisy and varied experimental data: the circadian clock in Arabidopsis thaliana. *Journal of theoretical biology*, 234(3), pp.383–393.

Makino, S. et al., 2000. Genes Encoding Pseudo-Response Regulators: Insight into His-to-Asp Phosphorelay and Circadian Rhythm in Arabidopsis thaliana. *Plant \& cell physiology*, 41(6), pp.791–803.

Martin-Tryon, E.L., Kreps, J.A. & Harmer, S.L., 2007. GIGANTEA acts in blue light signaling and has biochemically separable roles in circadian clock and flowering time regulation. *Plant physiology*, 143(1), pp.473–486.

Martinek, S. et al., 2001. A role for the segment polarity gene shaggy/GSK-3 in the Drosophila circadian clock. *Cell*, 105(6), pp.769–779.

Mas, P. et al., 2003. Dual role of TOC1 in the control of circadian and photomorphogenic responses in Arabidopsis. *The Plant cell*, 15(1), pp.223–236.

Matsushika, A. et al., 2000. Circadian Waves of Expression of the APRR1/TOC1 Family of Pseudo-Response Regulators in Arabidopsis thaliana: Insight into the Plant Circadian Clock. *Plant \& cell physiology*, 41(9), pp.1002–1012.

McClung, C., 2006. Plant circadian rhythms. *The Plant cell*, 18(4), pp.792–803.

McWatters, H. et al., 2000. The ELF3 zeitnehmer regulates light signalling to the circadian clock. *Nature*, 408(6813), pp.716–720.

Merrow, M.W., Garceau, N.Y. & Dunlap, J.C., 1997. Dissection of a circadian oscillation into discrete domains. *Proceedings of the National Academy of Sciences*, 94(8), pp.3877–3882.

Millar, A. et al., 1992. A novel circadian phenotype based on firefly luciferase expression in transgenic plants. *The Plant cell*, 4(9), pp.1075–1087.

Millar, A. et al., 1995. Circadian clock mutants in Arabidopsis identified by luciferase imaging. *Science (New York, N.Y.)*, 267(5201), pp.1161–1163.

Millar, A.J., 2015. The Intracellular Dynamics of Circadian Clocks Reach for the Light of Ecology and Evolution. *Annual review of plant biology*, 67(1), pp.annurev–arplant–043014–115619.

Millar, A.J. & Kay, S.A., 1991. Circadian Control of cab Gene Transcription and mRNA Accumulation in Arabidopsis. *The Plant cell*, 3(5), p.541.

Nagy, F., Kay, S.A. & Chua, N.H., 1988. A circadian clock regulates transcription of the wheat Cab-1 gene. *Genes \& development*, 2.

Nakamichi, N. et al., 2005. PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, together play essential roles close to the circadian clock of Arabidopsis thaliana. *Plant \& cell physiology*, 46(5), pp.686–698.

Narumi, R. et al., 2016. Mass spectrometry-based absolute quantification reveals rhythmic variation of mouse circadian clock proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24), pp.E3461–7.

Noordally, Z.B. & Millar, A.J., 2015. Clocks in algae. *Biochemistry*, 54(2), pp.171–183.

Nozue, K. et al., 2007. Rhythmic growth explained by coincidence between internal and external cues. *Nature*, 448(7151), pp.358–361.

Nusinow, D. et al., 2011. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

Pokhilko, A. et al., 2010. Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular systems biology*, 6, p.416.

Pokhilko, A. et al., 2012. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular systems biology*, 8, p.574.

Pokhilko, A., Mas, P. & Millar, A., 2013. Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. *BMC systems biology*, 7, p.23.

Rawat, R. et al., 2011. REVEILLE8 and PSEUDO-REPONSE REGULATOR5 form a negative feedback loop within the Arabidopsis circadian clock. G. P. Copenhaver, ed. *PLoS genetics*, 7(3), p.e1001350.

Salazar, J.D. et al., 2009. Prediction of photoperiodic regulators from quantitative gene circuit models. *Cell*, 139(6), pp.1170–1179.

Sanchez, S.E. & Kay, S.A., 2016. The Plant Circadian Clock: From a Simple Timekeeper to a Complex Developmental Manager. *Cold Spring Harbor Perspectives in Biology*, p.a027748.

Seaton, D.D. et al., 2015. Linked circadian outputs control elongation growth and flowering in response to photoperiod and temperature. *Molecular systems biology*, 11(1), pp.776–776.

Smith, K.M. et al., 2010. Transcription factors in light and circadian clock signaling networks revealed by genomewide mapping of direct targets for neurospora white collar complex. *Eukaryotic cell*, 9(10), pp.1549–1556.

Somers, D. et al., 2000. ZEITLUPE encodes a novel clock-associated PAS protein from Arabidopsis. *Cell*, 101(3), pp.319–329.

Song, Y. et al., 2012. FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. *Science (New York, N.Y.)*, 336(6084), pp.1045–1049.

Strayer, C. et al., 2000. Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. *Science (New York, N.Y.)*, 289(5480), pp.768–771.

Sun, L. et al., 1993. CA-1, a novel phosphoprotein, interacts with the promoter of the cab140 gene in Arabidopsis and is undetectable in det1 mutant seedlings. *The Plant cell*, 5(1), pp.109–121.

Takahashi, N. et al., 2015. A Hierarchical Multi-oscillator Network Orchestrates the Arabidopsis Circadian System. *Cell*, 163(1), pp.148–159.

Tseng, R. et al., 2014. Cooperative KaiA-KaiB-KaiC interactions affect KaiB/SasA competition in the circadian clock of cyanobacteria. *Journal of molecular biology*, 426(2), pp.389–402.

Voigt, C.A., Wolf, D.M. & Arkin, A.P., 2005. The Bacillus subtilis sin Operon: An Evolvable Network Motif. *Genetics*, 169(3), pp.1187–1202.

Wang, Z.Y. et al., 1997. A Myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene. *The Plant cell*, 9(4), pp.491–507.

Xie, Z. & Kulasiri, D., 2007. Modelling of circadian rhythms in Drosophila incorporating the interlocked PER/TIM and VRI/PDP1 feedback loops. *Journal of theoretical biology*, 245(2), pp.290–304.

Zeilinger, M. et al., 2006. A novel computational model of the circadian clock in Arabidopsis that incorporates PRR7 and PRR9. *Molecular systems biology*, 2, p.58.

Zeng, H. et al., 1996. A light-entrainment mechanism for the Drosophila circadian clock. *Nature*, 380(6570), pp.129–135.

Chapter 2
General Materials and methods

Computational tools and methods

All data analysis was in a MacBook Pro (Retina, 15-inch, Mid 2014), 2.5 GHz Intel core i7 16 GB 1600 MHz DDR3 NVIDIA GeForce GT 750M 2048 MB, otherwise stated. Python 2.7.12 and Ipython and Jupyter notebooks used for performing data, model analysis and prototyping of parameter fitting scripts. Critical python packages include scipy, numpy and matplotlib as general libraries for data and numerical analysis of models. All the scripts used for data analysis are in the attached CD arrange in folders for each chapter. For e.g. The Protein Binding Microarray analysis is in the Chapter 4 Chapter.

Model Fitting tools

Model fitting was performed using SloppyCell (python) which contains routines for parallelisation of parameter fitting based on PyPar 2.0 and open MPI. The Fitting algorithm of choice was Levenberg-Marquadt. Models were implemented in Antimony a human-readable analogue of SBML and converted into SBML Level 3 version 1 using Tellurium (python). Period was determined by calculating the time difference between two consecutive peaks. Peaks were obtained by performing a quadratic approximation to consecutive inflexion points. Scaling factors for models were determined automatically by

SloppyCell when waveform was fitted. When mass scale was fitted the SloppyCell scaling factors were set to 1. Scaling factors for parameters were introduced in models using Antimony. The nomenclature gmX (X = I, 9, 5, etc…) denominates transcript scaling factors for each of the genes in the clock model. While gpX denotes protein scaling factors.

Fitting considerations

SloppyCells exploits the use of an extended model were sensitivity equations are introduced nonetheless SloppyCell currently do not support sensitivity equations when period constraints are introduced. Instead a new SloppyCell Levenverg-Marquadt (LM) fitting function was created in order to support the use of period constraints. This function uses finite differences instead of sensitivity equations for determining the direction of steepest descent in the LM routine.

Experimental materials and methods

Media

LB (10g Tryptone, 5 g yeast extract, 10 g NaCl in 950 ml of ddH_2O)

ROBUST (1/2 MS salts pH 5.8 KOH adjusted without no sugars added)

Antibiotics stock solutions and working concentration

Kanamycin 50 mg/ml (1:1000), Spectinomycin 100 mg/ml (1:1000), Chloramphenicol 50mg/ml (34 μg/ml) Ampicillin 50 mg/ml (1:1000), Bilaphos (Sigma) 10 mg/ml (1:1000).

Gibson Assembly (Gibson et al. 2009)

Isothermal assembly, single-pot reaction for assembly of multiple overlapping DNA molecules by concerted action of T5 5' exonuclase, DNA polymerase and DNA ligase.

Reagents

Dithiotreital DTT: Sigma D9779 (prepare a 1M solution in ddH_2O)

NAD: Sigma N1511 (0.1M gentle heating at 45 ºC for to dissolve)

dNTPs: New England Biolabs N0446s

$MgCl_2$: 1M solution Sigma, M-1028

Tris-HCl pH 7.5: Invitrogen 15567-027

PEG-8000 (USB, 19959)

T5 Exo: Epicentre T5E4111K

Phusion Polymerase (NEB, F-530L)

Taq Ligase: NEB M0208L

Isothermal reaction buffer (ISO buffer) (25% PEG-8000, 500 mM Tris-HCl pH 7.5, 50 mM $MgCl_2$, 50 mM DTT, 1mM each 4 dNTPs, and 5 mM NAD) Aliquot in 40 µl and store at -20ºC

Assembly Master Mix

Reagent for 1 reaction	*Volume µl*
5x ISO buffer	4
T5 Exonuclase 10 U/µl	0.008
Phusion polymerase 2 U/µl	0.25
Taq Ligase 40 U/µl	2
ddH20	8.74

Aliquoted in 15 µl and store at -20ºC. The mixture can be used within 1 year of preparation. Assembly procedure:

1) Thaw a 15 µl aliquot and keep in ice until ready to use
2) Add 5 µl of DNA to assemble to the master mix. Fragments should be in equimolar amounts
3) Incubate at 50 ºC for 1 hour.
4) Transform E.coli (DH4alpha or ccdB resistant) (Life technologies)

Cloning recombinant version of NanoLUC

The NanoLUC sequence was amplified with Phusion DNA polymerase (New England Biolabs) and cloned into pUC54 with a synthetic 3xFLAG-10xHis (3F10H) tag (Eurogentech) using Gibson assembly. The sequence of Maltose Binding Protein (MBP) was amplified from plasmid pMJ806 (CAS9 containing vector (Jinek et al. 2012)) then amplified and fused to NL3F10H and pET28a

by Gibson assembly, sequence in CD Chapter 2. Constructs were transformed into DH5alpha *E. coli* and selected on Kan 50*μ*g/ml. Two independent transformants were verified by Sanger sequencing (Genepool, Edinburgh Genomics).

Purification of NanoLUC by Ni-NTA agarose

The resulting plasmid was introduced into BL21 DE3 Rossetta2 pLysS selected on 50*μ*g/ml 34*μ*g/ml LB. A single colony was selected and a starter culture prepared over night at 30ºC, in 5ml of LB broth supplemented with Kn50 and Cm50. Two cultures were prepared in 100 ml LB kan50 Cm34 in 250ml Erlenmeyer flasks, inoculated to an $O.D._{600nm}$ 0.01 and incubated at 30ºC with shaking at 200 rpm, and the O.D. monitored until it reached 0.5 $_{600nm}$. IPTG (Sigma) was added to a final concentration of 1mM and cultures were kept in the same growing conditions for another 8 hours. Cells were harvested by centrifugation at 4,000 rfcs, at 4ºC. The cells were suspended in 10 ml ice-cold Lysis buffer (50mM NaH_2PO_4, 300mM NaCl, 10mM Imidazole, pH 8.0 NaOH adjusted). Re-suspended cells were sonicated for 1 min with 10s on and 10 off and 50% amplitude in ice with a sonicator. The crude lysate was passed several times through a 25 syringe to reduce viscosity. 2 ml of lysate was centrifuged at 20,000 rfcs for 20 min at 4ºC, eliminating unbroken cells and debris. The supernatant transferred to a 1ml microfuge tube (safelock, Eppendorf), mixed with 250*μ*l of Ni-NTA agarose (Qiagen) and incubated at 4ºC for one hour with gentle agitation. Ni-NTA beads were recovered by

centrifugation at 20,000 rfc for 1 min and the supernatant removed. The agarose beads were then washed three times with Washing Buffer (50 mM NaH_2PO_4, 300mM NaCl, 20mM Imidazole, pH 8.0 NaOH adjusted), centrifuged 20,000 rfcs each time. Finally, MBP-NL-3F10 was eluted with Elution Buffer (50mM NaH_2PO_4, 300mM NaCl, 250mM Imidazole, pH 8.0 NaOH adjusted). The elution fraction was dialysed (10,000 MW cutoff) overnight at 4ºC in BI buffer (50 mM NaH_2PO_4, 300mM NaCl, 20mM pH 8.0 NaOH adjusted). The NanoLUC activity was determined from a $1x10^{-3}$ dilution in BI buffer. 20μl of diluted enzyme was mixed with 80 μl of BI buffer and 100 μl of 1:50 furimazine:NanoGlow assay buffer as a reference for starting activity

Cloning of genomic fragments of Arabidopsis clock genes.

All DNA purifications were quantified on NanoDrop (Thermo). All computational work for DNA design was carried out on Benchling (https://benchling.com/). The reference genomic sequence was TAIR10. Genomic DNA was extracted from 2-week-old Col-0 plants using the Qiagen plant extraction kit. The template was used to produce amplicons for Gateway cloning, in two rounds of PCR amplification using Phusion polymerase (New Englands Biolabs) (primers on table, synthesised by Eurogentec). Primers specific for the genomic regions of interest were used in the first round. They included an extended adaptor sequence to ease subsequent addition of attB and attP sites in a second PCR amplification. The resulting fragments were gel-purified with Qiagen gel purification kit and used for BP recombination with

the pDONR221 vector and BP clonase (Invitrogen). The resulting reactions were transformed into *E. coli* DH5alpha library efficiency cells (Life sciences). Two independent clones per gene were selected for future work. At this stage, the sequence was verified by Sanger sequencing (Genepool, Edinburgh Genomics), mapping sequencing results to TAIR10 in Benchling. The plasmids were then recombined into pGWB601:NL3F10H and pGWB635 by LR reaction using LR clonase (Invitrogen). The construction of pGWB601:NL3F10H is described in Chapter 3. The resulting reactions were transformed into DH5alpha *E. coli* and selected on Sp LB plates, then single colonies inoculated in 5ml LB incubated overnight with Sp and glycerol stocks with 25% glycerol final percentage flash frozen in liquid nitrogen and stored at -80ºC.

Table 2.1 Genomic regions amplified for LUC and NanoLUC translational fusions. Transcription starts site (TSS) +1 Stop codon eliminated, intron retained. The promoter sequences span from TSS of interest to the TSS of the upstream gene. Hence variability I starts sites.

Gene	*Start*	*End*
CCA1p:CCA1	-531	3043
LHYp:LHY	-899	3877
PRRp:PRR7	-1216	3692
TOC1p:TOC1	-1425	3110
LUXp:LUX	-614	1917
BOAp:BOA	-1002	1053

Table 2.2 Primers for genomic amplification and gateway cloning of clock genes

Gene	*primer*	*seq*	*Tm ºC*
CCA1	CCA1for	AGTCTTCTACCCTTCATGCATGGTTAGC	60.2
CCA1	attB1CCA1for	GTACAAAAAAGCAGGCTTAAGTCTTCTACCCTTCATGCATGGTTAGC	60.2
CCA1	CCA1rev	TGTGGAAGCTTGAGTTTCCAACCGC	61.9
CCA1	attB2CCA1rev	CAAGAAAGCTGGGTTTGTGGAAGCTTGAGTTTCCAACCGC	61.9
LHY	LHYfor	TTTTGGAATAATTTCGGTTATTTCAATTAGATTCGGGTAGT	60.3
LHY	attB1LHYfor	GTACAAAAAAGCAGGCTTATTTTGGAATAATTTCGGTTATTTCA	50.5
LHY	LHYrev	TGTAGAAGCTTCTCCTTCCAATCGAAGC	60.4
LHY	attB2LHYrev	CAAGAAAGCTGGGTTTGTAGAAGCTTCTCCTTCC	50
TOC1	TOC1for	GAGATCGCTCGGCTCAACAACAATATAT	59.1
TOC1	attB1TOC1for	GTACAAAAAAGCAGGCTTAGAGATCGCTCGGCTCAACAACAATATAT	59.1
TOC1	TOC1rev	AGTTCCCAAAGCATCATCCTGAGGAG	60.1
TOC1	attB2TOC1rev	CAAGAAAGCTGGGTTAGTTCCCAAAGCATCATCCTGAGGAG	60.1
PRR7	PRR7for	CGTCGGTGTAGAAGAGATCACGCTTAG	60.5
PRR7	attB1PRR7for	GTACAAAAAAGCAGGCTTACGTCGGTGTAGAAGAGATCACGCTTAG	60.5
PRR7	PRR7rev	GCTATCCTCAATGTTTTTTATGTCGTTATCATCAG	58.6
PRR7	attB2PRR7rev	CAAGAAAGCTGGGTTGCTATCCTCAATGTTTTTTATGTCGTTATCATCAG	58.6
LUX	LUXfor	AGAAGAAGATAACGTTTCGTCAGTTTGTGAAG	59.5
LUX	attB1LUXfor	GTACAAAAAAGCAGGCTTAAGAAGAAGATAACGTTTCGTC	50.2
LUX	LUXrev	CATGATACTTTGTATGATCCTCTCCTGAACAGATG	60
LUX	attB2LUXrev	CAAGAAAGCTGGGTTCATGATACTTTGTATGATCCTCTC	50.7
BOA	BOAfor	TGTTTCAAAGGCAGCGTCTGCGTTATG	62.3
BOA	attB1BOAfor	GTACAAAAAAGCAGGCTTATGTTTCAAAGGCAGCGTCTG	55.8
BOA	BOArev	CTTGTCATTCACGTCACCACCACCAAC	62
BOA	attB2BOArev	CAAGAAAGCTGGGTTCTTGTCATTCACGTCACCACC	55.8

Protoplast isolation, transformation and imaging.

Modification from the base protocol are indicated in this section (Hansen & van Ooijen 2016). Plants were grown in 12L:12D photoperiod under 120 μmoles/cm^{2}s^2 cool white fluorescent light for four weeks. Protoplast isolation started at ZT2 using 0.5% CELLULYSIN cellulase from *Thrichoderma viridae* 100 kU *(Sigma) and* 0.25% Pectinase from *Rhizopus sp* 5 kU (Sigma). Vectors

with constructs of interest were purified with Qiagen Maxi Kit. These preparations were then used for transforming protoplasts by PEG 4000 with 4.6 pmoles of plasmid concentrations described by Hansen et al. Imaging solution for NanoLUC contained 1:50 Furmiazine:Imaging Buffer, instead of luciferin..

Plant transformation and selection of primary transformants

The vector DNAs were prepared from *E. coli* DH5alpha cultures using Qiagen miniprep kit. 100 *µ*g of DNA was transformed by a liquid nitrogen freezing method (Wang 2006) into *Agrobacterium tumefaciens* ABI, kindly donated by Professor Seth Davis (The University of York). Individual colonies were taken for preparing a starter culture grown at 30ºC 200 rpm over-night using LB with antibiotic selection Kan, Cm, Sp. Then 1 ml of starting culture inoculated in 50 ml LB media 50 *µ*g/ml Kan, 100 *µ*g/ml Sp, 3.4 *µ*g/ml Cm, 150 *µ*M Actetosyringone (Sigma) and grown for 24h at 28ºC 150 rpms. *A. tumefaciens* cells carrying plasmids of interests were harvested by centrifugation at 5000 rpm for 15 min at 28C and re-suspended with infiltration media (1/2 MS medium containing 5% Sucrose, 3mM MES, pH 5.5 KOH adjusted, 200 *µ*l/L Silwet L-77, 150*µ*M Acetosyringone).

Variability in flowering times on mutants required different number of plants for the transformation. All plants in each genotype were clipped for promoting

branching. The secondary inflorescences were dipped in infiltration media containing Agrobacterium with the desired construct. Plants were incubated in the dark for 24 hours at 21ºC. The plants were grown to maturity, seeds harvested and stored for further analysis resulting in T0 seed generation. These seeds were then surface sterilised, stratified at 4ºC in ddH20 for 48h and sown on F2+sand compost. Seedlings were germinated in 16L:8D photoperiod 21ºC 100 μmol/s^{2}cm^2 white light. Plants were treated with 2.5 mg/L BASTA, 0.01% Triton X-100, applied every two days. Resistant plants could be appreciated after one week of selection. The herbicide treatment continued to avoid the growth of sensitive late germinators, but with longer time interval between treatments. Seed from individual plants was collected resulting in T1 generation seed, which was stored for segregation analysis (Figure 5.2).

Selection in the T1 generation

Segregation analysis was performed in 10 cm Petri dishes containing 15 ml of ROBUST media (1/2 MS salts without sucrose pH 5.8 adjusted with KOH) supplemented with 10 mg/L Bialaphos (Sigma) (Figure 2.2). For robust determination of segregation ratios, 40 seeds per plate per line were stratified in cold for 48 ºC and the transferred to 16L:8D 100 μmoles/s^{2}cm^2 21ºC. Lines presenting segregation ratios of 3:1 resistant:sensitive were selected for further analysis. In general, around 30-40% of primary transformation lines

presented a 3:1 ratio. From this analysis, 8 plants per line were transferred to soil and T2 generation seeds were collected from individual plants.

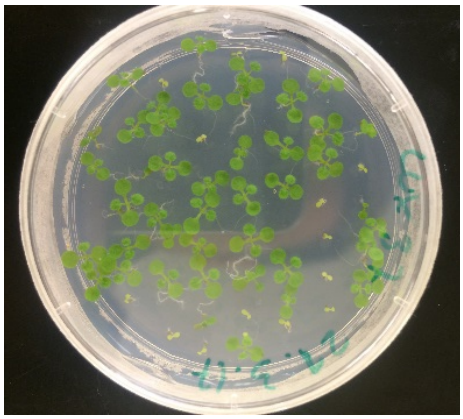


Figure 2.1 Segregation analysis with Bialaphos in ROBUST media. Shown is a T2 LUXNL37.2 segregating population. Herbicide-sensitive plants can be observed after 2 weeks in long day conditions. 8 resistant plants were transferred to soil for further analysis.

Selection of T2 homozygous lines

The number of lines generated in this work required a different approach for selecting homozygous segregants, because more than 1,200 seed batches required segregation analysis for selecting T2 seed batches that homozygous. I exploited the advantage of selecting with BASTA on soil (Figure 2.3). Seed batches that were a 100% resistant were selected as homozygous candidates for analysis of period complementation in constant light by luciferase imaging assay.

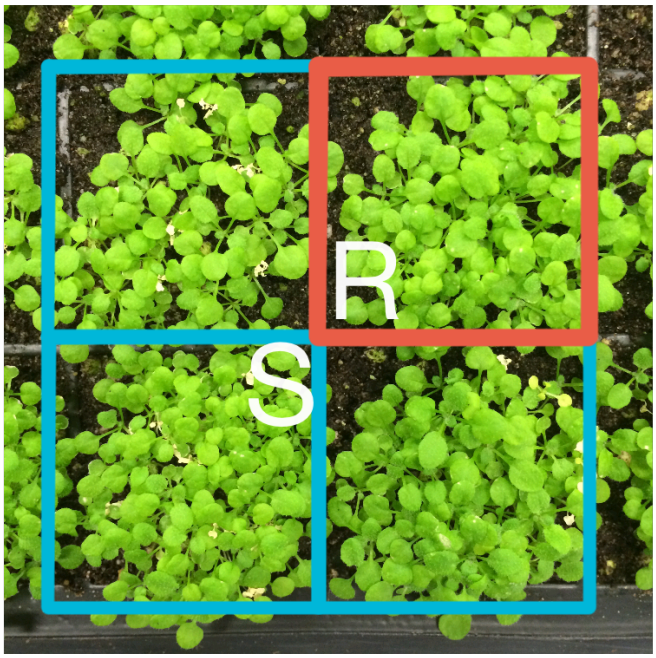


Figure 2.2 T3 selection on soil using BASTA. Progeny of T2 plants Col-0 prr7/9 PRR7p:PRR7-NL3F10H on soil. S) batches that present with dead sensitive plants. R) resistant batch

Luciferase imaging assay

Seeds were surface sterilised 20% bleach 0.01% Triton X-100. Between 10-20 seeds were deposited on hand-made wells using 1.5 ml microfuge tubes, in square tissue culture dishes (Figure 2.4). The tubes were immobilised using ROBUST media (0.5x MS agar, pH 5.8) as plant substrate (Figure 5.3). The seeds were stratified at 4ºC for 48h, then transferred to 21ºC 50 μmol of monochromatic blue and red light in 12L:12D photoperiod for 7 days, in a Percival growth chamber. Seedlings were sprayed with 5 mM luciferin 0.01% Triton X-100 on the 7th day at ZT 2 and returned to entrainment conditions for another day before starting imaging. The ORCA2 EM-CCD -75ºC cooled

camera took images every two hours with an exposure time of 30 min, essentially as described (Edwards et al. 2010).

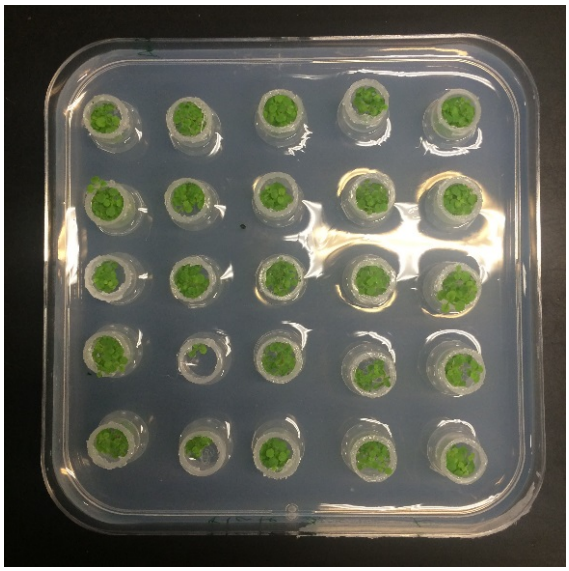


Figure 2.4 Plate set up for imaging experiments. 25 wells handmade can be fit in square plates. ROBUST media used as substrate.

Period determination

I determined the signal intensity from images using ImageJ64. Regions of interest for each seedling group were selected, luminescent signal was quantified, and background subtracted. The resulting time-series were then analysed in the online Biodare2 resource using the FFT-NLS algorithm(Plautz et al. 1997). The period was determined by considering time-points starting from the second day in LL. Biodare2 results were exported as spreadsheet files and analysed using *ad hoc* python scripts.

Determination of NanoLUC activity in plant extracts.

Samples of plants were collected in pre-weighed 2 ml microfuge tubes (safelock, Eppendorf) with 5 mm stainless steel grinding balls, and flash frozen in liquid nitrogen. The tissue was ground twice at 30Hz for 1 min in a Tissue Lyser (Qiagen). The samples were flash frozen between grinding steps, then placed on ice and 150 μl of BSII buffer was added to protect the samples from proteolysis, without phosphatase inhibitors (Huang et al. 2016). The tube is weighed and further BSII buffer added to adjust tissue concentration to 0.4 gFw/ml, and centrifuged at 10,000xg for 10 min. For NanoLUC assays, 20 μl of clarified plant extract was mixed with 80 μl of buffer BI (50 mM NaH_2PO_4, 300 mM NaCl, pH 8.0 NaOH adjusted). Samples were loaded in a 96-wel black flatl plate, mixed with 100μl of 1:50 Furimazine:NanoGlow buffer (Promega) and incubated at 21^{o}C for 10 min. Luminescence signal was determined in a Berthold Tristar plate reader, with a 1.5s integration time. Assay plates could subsequently be visualised by low-light imaging to give a qualitative overview of the results (Figure 2.5).

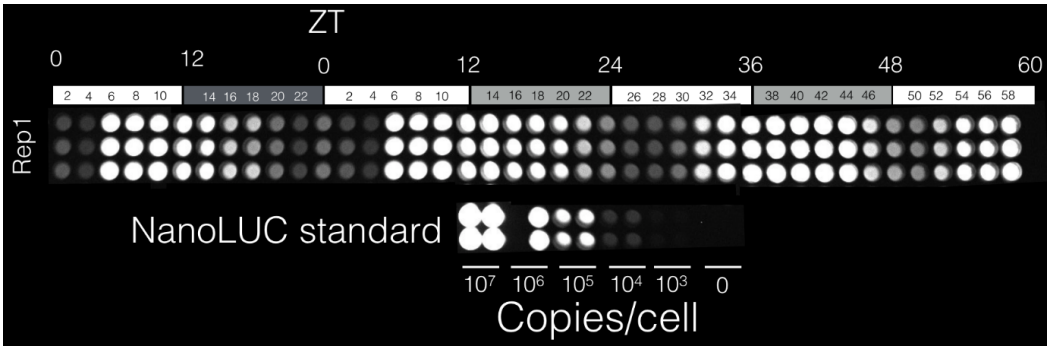


Figure 2.5 Quantification of PRR7-NanoLUC activity of plant extracts in a in 96-well plates. One of two biological replicates shown, with three technical replicates. The calibration curve is placed in the same plates to control for temperature and incubation time. Images of the assay plates were captured in an ORCA-II EM-CCD cooled camera, exposure time 2 minutes, and re-assembled to give a montage in time order. Light grey bar, predicted dark interval; dark grey bar, dark interval; white bar, light interval. Time in hours.

Construction of calibration curve for absolute quantification in plant extracts.

Plates inoculated with Col-0 seed were grown under the same photoperiod conditions to the plants to be analysed. Plant tissue was harvested as described, making aliquots of 0.1 gFW. MBP-NL3F10H protein was prepared by the method described above. and then quantified by the linearized Bradford assay protocol using both Bovine serum albumin BSA and Ovoalbumin as standards (Ernst & Zor 2010). Then aliquots spiked with purified enzyme to generate a curve with: 0, $1x10^2$, $1x10^3$, $1x10^4$, $1x10^5$ and $1x10^6$ molecules per cell, assuming $25x10^6$ cell/gFw in Col-0 (Flis et al. 2015). Absolute quantities of protein then were determined by log-transforming the intensity, performing a linear regression and inverting the linear model (Figure 2.6).

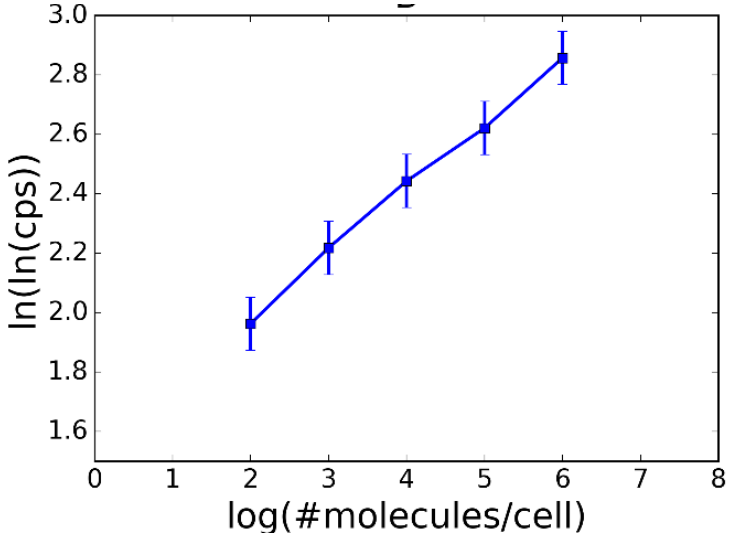


Figure 2.6 Calibration curve generate by spiking Col-0 tissue aliquots with purified MBP-NL-3F10H, assuming 25x10^6 cells/gFw.

Plate reader bioluminescence time series in vivo

2-week-old plants were grown under 12L:12D photoperiod 140 μmol/cm^{2}s^2 white light. Individual plants were transferred to each well of a 96-well plate, containing 100 μl agar ROBUST media per well and 100 μl of 1:50 furimazine:0.01% Triton X-100. Then transferred to plate reader conditions, 50 μmol/cm^{2}s^2 red and blue LED light. Luminescence measurements were taken every 30 min, with one-minute dark delay before performing 1.5s signal integration.

Edwards, K. et al., 2010. Quantitative analysis of regulatory flexibility under changing environmental conditions. *Molecular systems biology*, 6, p.424.

Ernst, O. & Zor, T., 2010. Linearization of the bradford protein assay. *JoVE (Journal of Visualized Experiments)*, (38), pp.e1918–e1918.

Flis, A. et al., 2015. Defining the robust behaviour of the plant clock gene circuit with absolute RNA timeseries and open infrastructure. *Open biology*, 5(10), p.150042.

Gibson, D.G. et al., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods*, 6(5), pp.343–345.

Hansen, L.L. & van Ooijen, G., 2016. Rapid Analysis of Circadian Phenotypes in Arabidopsis Protoplasts Transfected with a Luminescent Clock Reporter. *JoVE (Journal of Visualized Experiments)*, (115), pp.e54586–e54586.

Huang, H. et al., 2016. Identification of Evening Complex Associated Proteins in Arabidopsis by Affinity Purification and Mass Spectrometry. *Molecular & cellular proteomics : MCP*, 15(1), pp.201–217.

Jinek, M. et al., 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, 337(6096), pp.816–821.

Plautz, J.D. et al., 1997. Quantitative analysis of Drosophila period gene transcription in living animals. *Journal of Biological Rhythms*, 12(3), pp.204–217.

Wang, K., 2006. *Agrobacterium Protocols*, New Jersey: Springer Science & Business Media.

Chapter 3
Incorporating absolute units into the transcriptional component of the circadian clock model of Arabidopsis

In this Chapter, I Introduce absolute units at the transcript level for the first time in Arabidopsis clock models. First, I describe the starting data set used in the analysis. Then, I do a short revision of the mathematical formalisms under consideration. Also, I performed three important modifications in the way the model is approached. 1) Elimination of the quasi-steady state assumption for the ELF3:ELF4 dimer formation in P2011. 2) Assignment of Lmod activity to RVE8 protein, this modifies the way the *lhy/cca1* double mutant is simulated. 3) Generating a revised PRRs regulation based on transcriptional repression which can be considered to be an intermediated step between P2011 and F2014. Then, I discuss the type of experimental error observed in the considered data. I also, show how the period constrains should be introduced in the likelihood function used for parameter inference. After this short section, I present the fitting results of U2017.1 and U2017.2 to the RNA data in absolute unit. Finally, I compare the inferred transcription rates to genome wide distribution of transcription rates which allow me identify clock elements for which transcription rates are unrealistic. The use of absolute units allows new

ways for testing mathematical models. In particular testing if transcription rates are biologically meaningful. However just rescaling models to match the mass scale is not enough for verifying if transcription rates are biologically valid. Here, I show that by integrating qRT-PCR and microarray data from the literature, it is possible to build an empirical distribution for transcription rates that presents absolute units. This distribution provides a valid framework for evaluating transcription rates in clock models. This shows that the use of absolute units allows for the first-time testing aspects of the model that were not accessible before.

Presentation of TiMet data.

Modelling efforts in Arabidopsis up to now have been based on data in arbitrary units. This is the case for both transcript and proteins levels. Generally, the data is normalised against an internal standard that present should present constant levels across the day, tissue and experimental conditions. Nonetheless, in the context of the TiMet project in order to avoid unexpected artefacts associated with normalisation, our collaborators performed absolute quantification of RNA levels (Flis et al. 2015). This helps in circumventing the lack of standardisation in RNA time series that comprise interpretation of model fitting results (Fogelmark & Troein 2014). Two ways of impacting estimations can be foreseen in the normalisation process. First, if the internal standard used in the normalisation process has some sort of oscillatory dynamics or trends this will modify the original wave RNA wave form resulting

in parameter inference artefacts. A less appreciated issue associated with normalisation is that the uncertainty for an experimental point will be dependent on the variability internal standard in a multiplicative way. This might result in non-structural parameter identifiability issues that could have been avoided by normalising the RNA levels to a more stable observable like grams of fresh weight.

The starting data provided by our collaborators consists in time-series of RNA levels in units [transcripts]/[cell] (Figure 3.1). The time series correspond to the clock genes represented in the P2011 model. The data includes the genotypes: Col-0, *lhycca1, prr79, toc1* and *gi*. Plants were grown for 20 days in a 12L:12D photoperiod. On day 21, sampling started in intervals of 2 hours for 72 h. After one cycle of 12L:12D, plants were transferred to constant light for 2 days in order to capture the relaxation of the oscillator after entrainment conditions. The levels were determined by qRT-PCR. The absolute scale was set by spiking tissue with known concentrations of standard RNAs. Therefore, calibrating the RNA levels in units [transcripts]/[gFw]. Cell number per gFW was determined by quantifying the genome copy number per gFw using artificial DNA spikes and using as internal standard genes that are known to be present in a single copy in the genome. The ploidy was estimated using flow cytometry. Resulting in estimated of $25x10^6$ cell/gFw (Flis et al. 2015).

The mathematical framework.

The staring clock mathematical model is P2011 (Figure 3.2). The made of systems of coupled non-linear ordinary differential equations (ODEs). Mass enters the system at a rate *n* from the cellular metabolic network that is not modelled. The rate *n* (transcription) in each gene can be modulated by the presence of transcription factors (TFs), formally expressed with the following ODE

$$\frac{dcT_m}{dt} = nf(cN_1, cN_2, \ldots, cN_n) - mcT_m$$

Equation 3.1

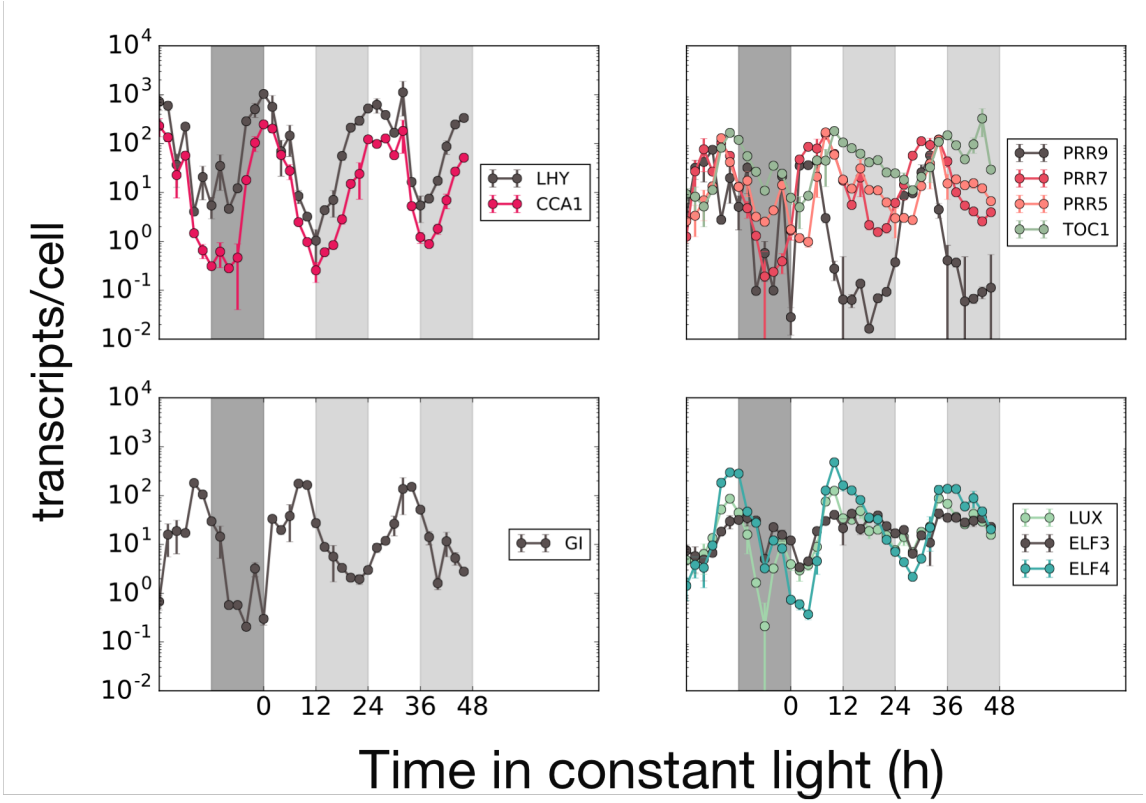


Figure 3.1 Col-0 qRT-PCR with absolute scale on the mass axis for genes in P2011 model (Figure 3.2). Balls: mean values error bara: S.E.M. Dark grey represents night; light grey represent expected external night in constant light conditions.

Where cT_m is the mRNA concentration of the TF. $f(cN_1,cN_2,\ldots,cN_n)$ is a functional form that modulates *n*. This is most of the coupling in the equations happens. The instability of the mRNA is captured by the decay constant *m* which can be modified by light in some cases.

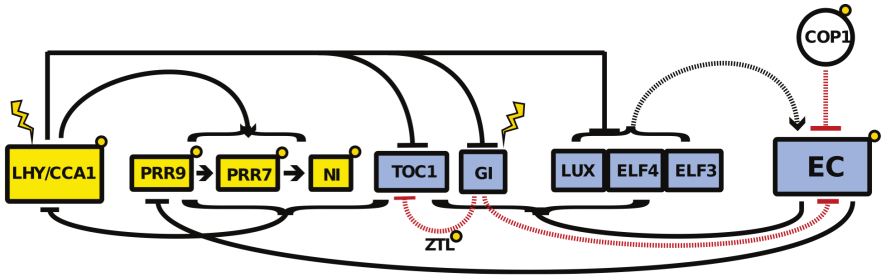


Figure 3.2. P2011 netwrok architecture. Morning loop elements (yellow). Evening loop elements (blue). Solid arrows represent transcriptional regulation, while dashed ones denote post-translational regulation. Direct light inputs to the clock are depicted with bolts and indirect ones with small yellow circles.

For a single transcriptional repressor, their action on a particular promoter is modelled by the Hill function

$$f(cT) = \frac{g^h}{g^h + cT^h}$$

Equation 3.2

Where the hill coefficient *h* generally can integer values that represent the degree of cooperativity for example *h=2* represents dimerization. The affinity

of a TF for a promoter is captured by the dissociation constant *g*, currently in arbitrary units.

The mRNA is translated into protein at a rate

$$\frac{dcT}{dt} = pcT - bcT$$

Equation 3. 3

p denotes the mRNA translation rate and *b* represents the protein decay constant. The prefix *c* in the variables denotes concentration in order to avoid the extensive use of brackets in the notation. Only the decay parameters present units that provide real time scale in hours (h). Figure 3.3 exemplifies the result of numerically solving P2011 for cLm (*CCA1/LHY* transcript) and cT (TOC1 protein). Note the lack of absolute units on the y-axis.

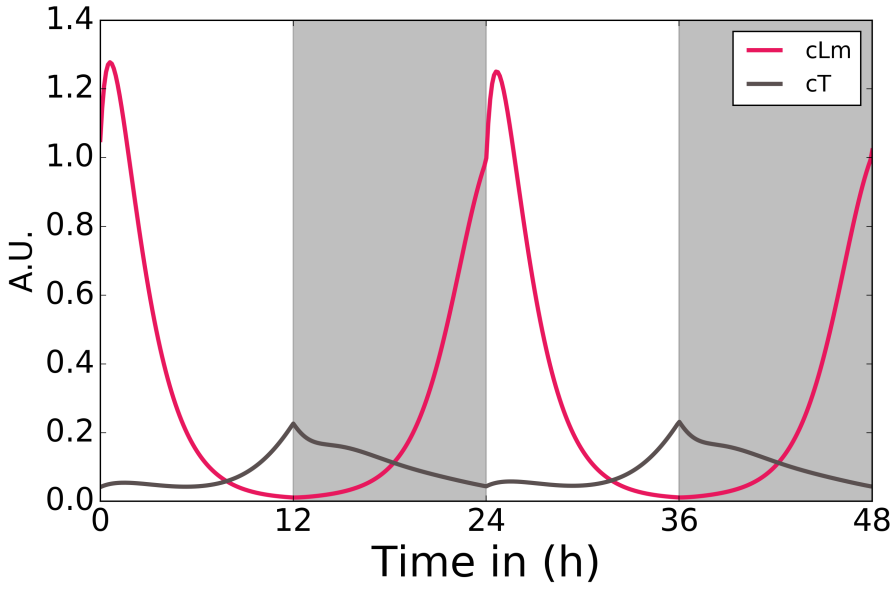


Figure 3.3 Exemplary predictions generated by P2011. cLm denotes CCA1/LHY transcript. cT, TOC1 protein. cT is a repressof of cLm. Grey bars represent night

Justifying the use of P2011 as the starting model

As described in Chapter 1, currently there are several mechanistic models that describe the transcript relative abundance for key clock genes (Pokhilko et al. 2012; De Caluwé et al. 2016; Foo et al. 2016). Each of these models have different aims and assumptions. The focus of this work is introducing absolute units in plant clock models for the first time. Therefore, the use of a well-established model like P2011 should facilitate the interpretation of results after introducing the absolute units. Also, it is important at this point to project into the future what the use of the model in absolute units might be. This is

important as it will provide a selection criterion for performing work on a model that will be used in plant systems biology.

P2011 is a favourable candidate because it has been useful for interpreting the impact of the clock on plant physiology using a multi-scale modelling approach.(Seaton et al. 2015; Chew et al. 2017). Therefore, if the model will be used for understanding clock outputs, some degree of detail is required for example compared a more compact representation for the clock. For example, C2016 model leaves out GIGANTEA (GI) and ZTEITLUPE (ZTL) (De Caluwé et al. 2016). These genes are important regulators of stability of the PRRs proteins which also present and effect on the stability of CONSTANS (CO) protein a crucial regulator of flowering time (Hayama et al. 2017).

An alternative to this using a model that is too simple, right at the start of this project F2014 model was published. F2014 has incorporated new variables and a complete new rewiring for how activation happens in the oscillator. Many of the new variables represent pools of protein in different compartments for which experimental measurements have not been carried. Measurement of these pools might remain open possible for several years, even though they have been observed experimentally. F2014 could be the successor of P2011 but plenty of work needs to be carried out in order to characterise the proposed architecture. One example being understanding what the dominant parameters are that determine its dynamics. While some solid work has been

done in this direction for P2011 (Schmal et al. 2013; Domijan & Rand 2015; Domijan et al. 2016) A drawback for F2014 is that the number of parameters have kept artificially low by setting mRNA translation to 1. This feature is problematic because I will introduce absolute units for the protein component in later chapters. New connections in the model do increase the complexity of the model and they should be considered while preforming model selection (Keily et al. 2013).

Starting with P2011 I will suggest some changes in how the network can be interpreted. This work touches, the Evening Complex (EC), the Lmod variable and the regulation of the PRRs. The EC and Lmod new treatments are relevant in future chapters. However, the PRRs revision is intended to understand if the new proposed regulation by other groups leads into better fits starting from P2011. Therefore, two new models are introduced in the next sections U2017.1 and U2017.2.

Results

Elimination of the quasi steady-state assumption for the Evening Complex, recovering P2011 dynamics and a role for *BOA*.

In P2011 an intermediary complexes ELF3:ELF4 forms which then interacts with LUX in order to form the Evening Complex (EC). These complex in turn represses the expression of *PRR9, TOC1, LUX and ELF4.* ELF3:ELF3 dimer

formation can be assumed to work at a faster time scale relative to transcriptional and translational processes. This can help in reducing the number of ODEs by setting the fast equilibrating process to zero. The parameters associated with these equations should be consistent with this assumed time-scale. However, inspecting the algebraic equations for the dimerization of ELF3:ELF4 reveals that the complex never dissociates, resulting in an infinite affinity. This assumption combined with the published parameter set of P2011 results in slow formation of the EC and levels peaking at the middle of the night (Figure 3.5 A). I am interested in removing unrealistic assumptions from the model therefore I re-introduced the ODEs for variables that are in quasi-steady state in P2011. This new model takes the name U2017.1 according to the plant circadian model nomenclature (Flis et al. 2015).

In P2011 the slow formation of the EC is an important characteristic. It allows short period oscillations in the *lhy/cca1* simulated mutant (Figure 3.5B). Slow formation therefore allows enough delay between the synthesis of the EC components and the EC formation which in turn feedbacks and repressed the expression of its components (*LUX, ELF3 and ELF4*).

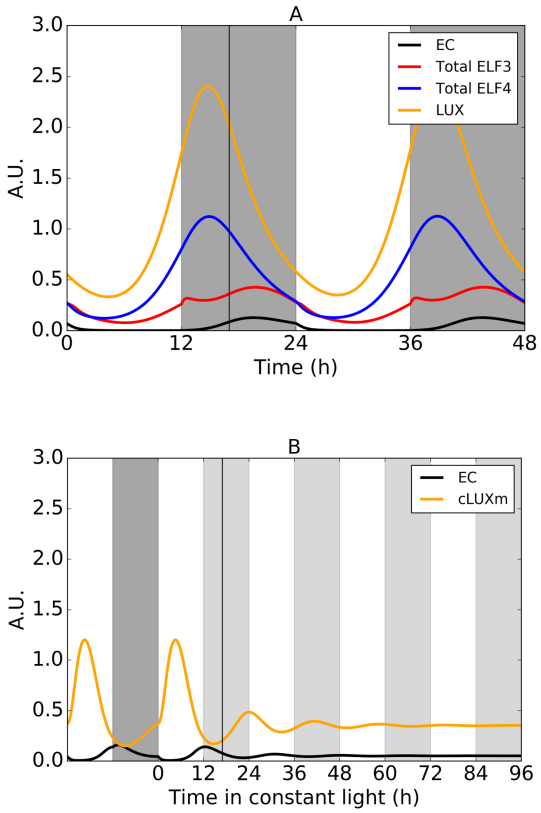


Figure 3.4 Dynamics of EC species in WT and lhy/cca1 mutant in P2011.1. A) simulation of WT behavior in LD cycles for the proteins conforming the EC, vertical black line represents midnight. The total pool of ELF3 is plotted in red. B) Simulation of LUX mRNA in the lhy/cca1 in P2011.1 double mutant is in antiphase to the EC which negatively feedbacks into LUX promoter. Model was entrained for 10x12L:12D cycles for obtaining initial conditions.

The complexity of the model results in a vast parameter space. This complicates the fitting process. I decided to recover P2011 dynamics by fitting U2017.1 to perfect data generated from P2011. In particular I simulated data for the variables and networks present in TiMet data. Nonetheless, I also included the protein pools of the EC intermediaries. All parameters in U2017.1 were allowed to be optimisable. The optimisation procedure and tool used are described in the methods Chapter.

U2017.1 closely matches the dynamics of the P2011 EC (Figure 3.8). A summary for the cost per variable shows that this is the case for the rest of the variables used in the fitting (Figure 3.9). From here onwards I use U2017 for all simulations and analysis unless otherwise specified.

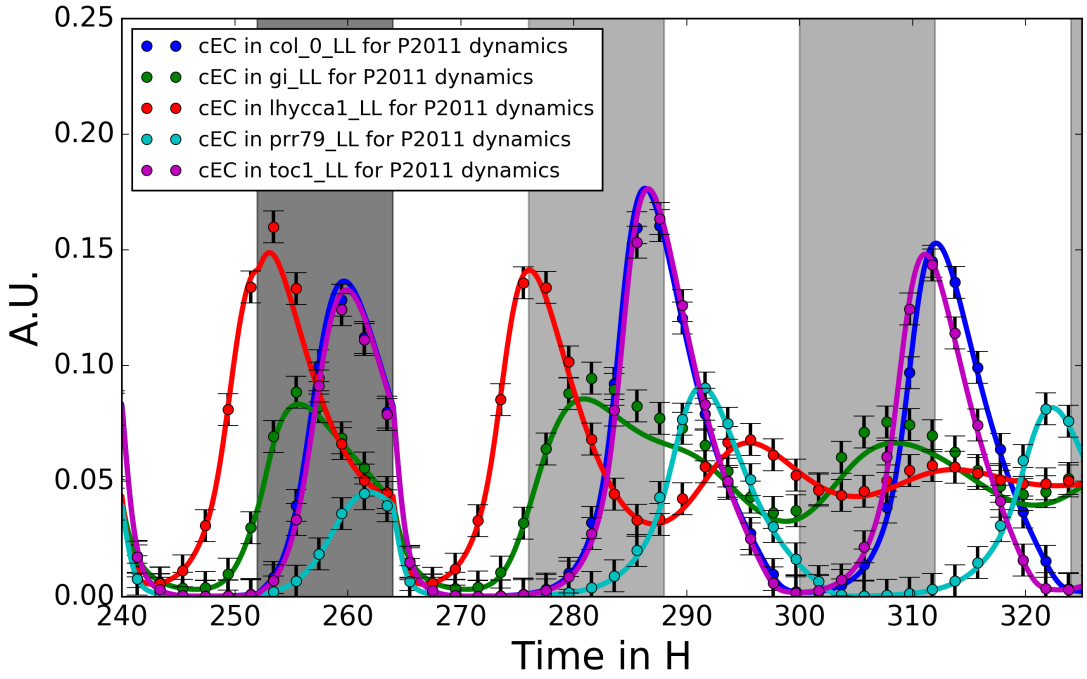


Figure 3.5 U2017.1 recovered the dynamics of P2011. The mutants present in TiMet data were used in the fitting process. Balls, perfect data from P2011. Solid lines U2017.1 fitted to P2011 perfect data.

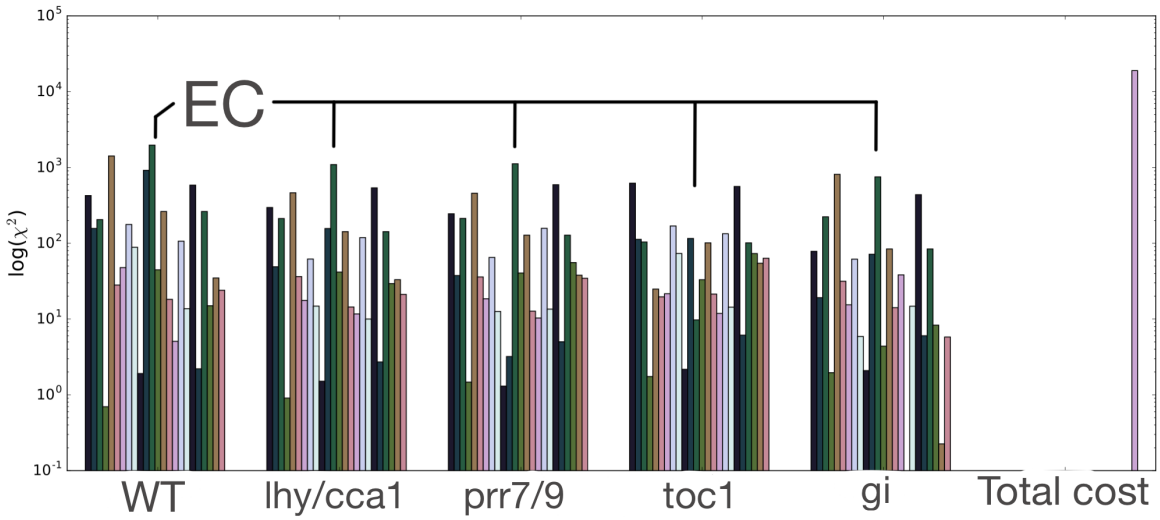


Figure 3.6 Summary of goodness of fit of U2017 to P2011 dynamics for mutants that present in TiMet data. The Evening Complex costs are highlighted. Total cost is the sum of all costs.

BROTHER OF LUX ARRHYTHMO (BOA)

Western-blot data suggest that the EC presents highest detected levels at dusk in a 12L:12D photoperiod (Nusinow et al. 2011). This is more than 6 h before U2017.1 model predictions (Figure 3.6). It is possible that the EC in U2017.1 represents a required repressive activity at the middle of the night rather which might be fulfilled partially by the EC in the plant in the second half of the night. The complementary activity might be mapped to *BROTHER OF LUX ARRYHTMO (BOA)* also known as *NOX. BOA* RNA profile is rhythmic with peaks at ZT16 in LD and ZT12 in LL. CCA1 protein binds to the promoter of BOA in a repressive manner. BOA in turns binds to the promoter of CCA1 and to synthetic multimers of the LUX binding motif in yeast one-hybrid experiments. Over-expression results in la longer period phenotype (26.85 s.d.=0.98 h vs 24.95h s.d.=0.04). Interestingly the transcript remains rhythmic

in the *lhy-1/cca1-11* double mutant in LL with expression across the day. Unfortunately, *boa-1* T-DNA line did not supress RNA levels neither RNAi targeting *BOA* transcript result a significant impact on the period of the clock (Dai et al. 2011; Helfer et al. 2011). The western blot time-series for BOA protein using antibodies against the native protein suggest a very low amplitude oscillation at subjective night. I decided to obtain additional evidence for BOA protein by generating a translational fusion of BOA with the firefly luciferase gene (LUC) and study its behaviour in stable transgenic plants. I super-transformed Col-0 plants by flora dipping as described in the methods chapter. Then, using a segregating population of Col-0 BOAp:BOA-LUC, I performed a plate reader experiment using 2 week old plants (Figure 3.10).

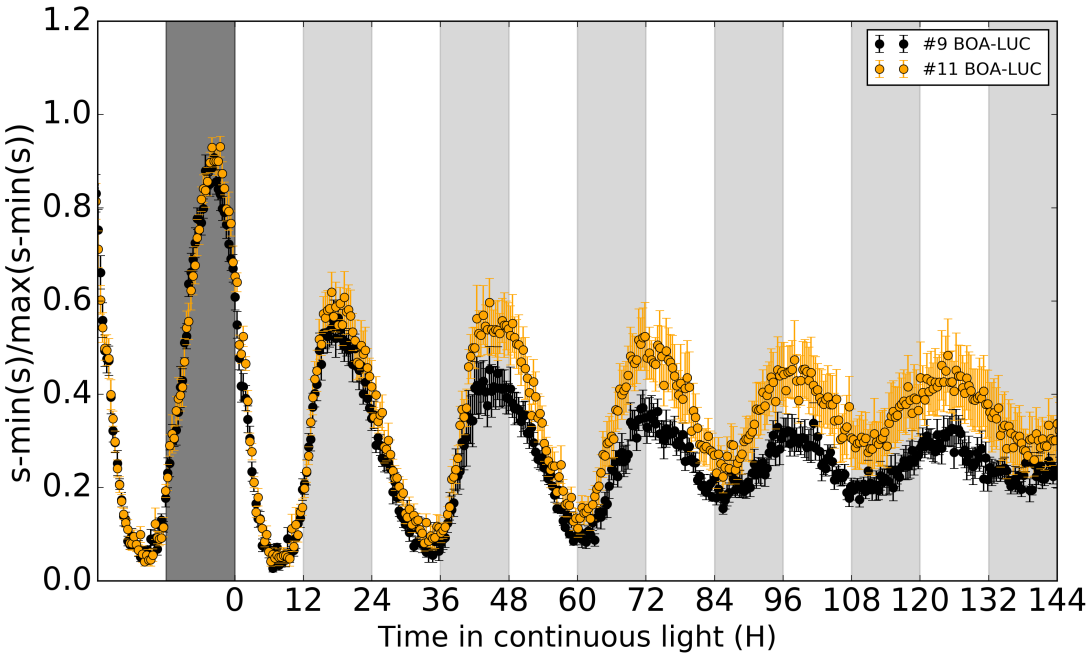


Figure 3.7 Dynamics of BOA-LUC in Col-0. Plants grown for two weeks in ROBUST media supplemented with 10mg/L Bilaphos were transferred to 96 well plats. Plants were then allowed to adjust for two days (2x12L:12D) to the

new light conditions (50 µmol/cm^{2}s^s) before measurements using a Tristar plate reader took place. N=4 error bars: S.E.M.

The time-series show that BOA-LUC signal (s) has peaking levels at a similar phase as the EC simulations from U2017.1 in LD. This experimental result provides evidence for having BOA protein in mind if work in the direction of moving EC:LUX version to its correct phase in future models. This has been proposed by F2014 nonetheless I want to remark that this model does not recapitulate the *lhy/cca1* double mutant therefore a more careful analysis and modelling work needs to be carried out if the EC variants are the key players in that mutant. I will present more work on this direction in future chapters.

The modified LHY/CCA1 (Lmod) and a role for RVE's.

The first mutant modelled in detail is the *lhy/cca1*.

This are interpreted mathematically as changes in model parameters. Mutations that result in phenocopies can occur at different levels in the model. For example, a null mutant can be made at the transcriptional level or at the translational levels by letting of the transcription rate and translation rate to 0. One of the first mutants modelled was the *lhy/cca1* double mutant in clock models. I suggest a new representation for this mutant in U2017.1.

The P2010 model introduced the variable L_{mod}, which is used to activate the *PRR7* at the middle of the afternoon. Phosphorylation of CCA1 by CK2 was proposed as an exemplary mechanism for the creation of this species {Daniel:2004mx, Pokhilko:2010vw}. However, LHY/CCA1-LIKE family of transcription factors include several genes expressed at down. Among them RVE8 has the most important effect on the clock. The *rve8* mutant has a possible long period or possibly a phase delay. The overexpressor line *RVE-OX* presents a phase advance or short period (Figure 3.9). Experimental data suggest that it acts as an activator of *PRR9, PRR5, TOC1* and *LUX* by binding to the Evening Element (EE) present in the promoter of these genes {Farinas:2011rt,Farinas:2011rja (Hsu & Harmer 2012). One of the most striking features of RVE8 protein is that it presents a peak of expression 6 hours after the peak of its transcript, observed by western blot. This feature is crucial because it suggest that RVE8 protein can be mapped to Lmod (Figure 3.8).

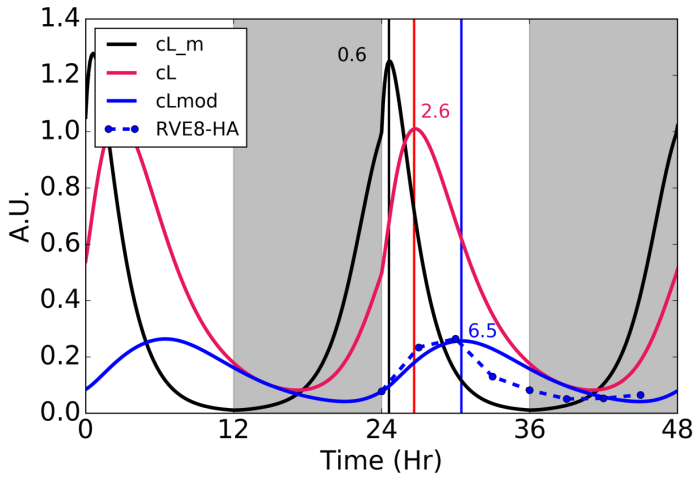


Figure 3.8 RVE8-HA profile matches Lmod profile in U2017.1. cL_m transcript representing CCA1/LHY, cL LHY/CCA1 protein. cLmod, modified originally a proposed modified version of CCA1/LHY protein. Western blot data of RVE8-HA from (Rawat et al. 2011).

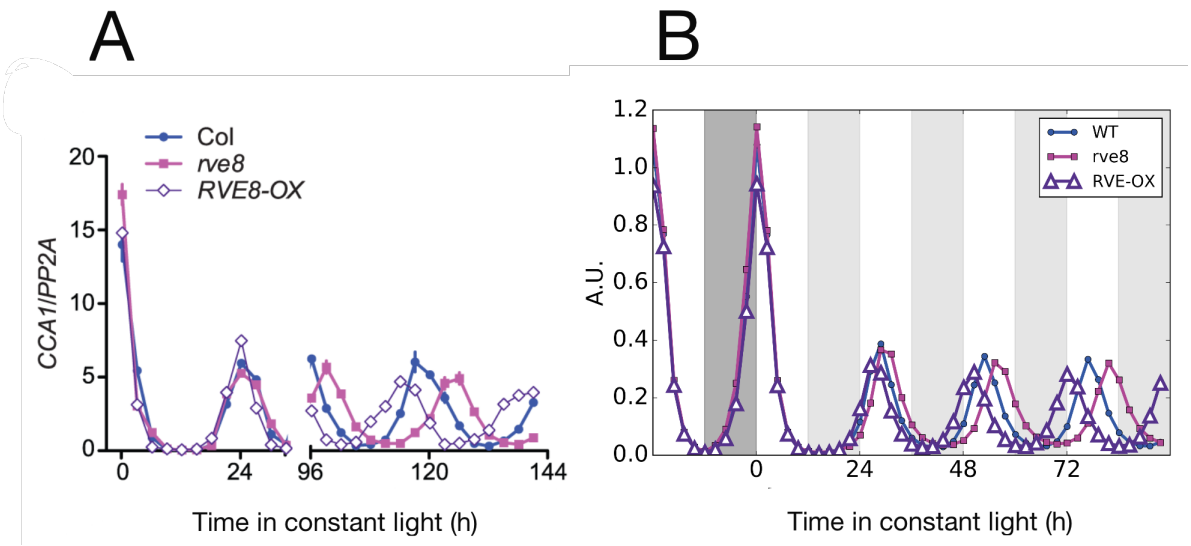


Figure 3.9 Emulating the effect of RVE8 in U2017.1. A) qRT-PCR data for CCA1 in Col-0, rve8 and RVE8-OX from (Rawat et al. 2011). U2017.1 can emulate the phenotype of RVE perturbations by assuming a complex formation at ZT6. Dark grey night, light grey expected night in constant light conditions.

Emulating the genetic perturbations in the U2017 model is a subtle process, which might explain why this mapping was originally overlooked. In U2017 the *rve8* mutant corresponds to setting to 0 the cL (LHY/CCA1) to cLmod transformation rate. This result in a long period phenotype in the model (Figure 3.9 B rve8 trace).

Emulating the over-expressor is difficult because an assumption is required to produce the expected effect. It is possible that the Lmod variable represents a complex formation that peaks at ZT6.5. Therefore, in plants RVE-OX there will more RVE8 protein available for early formation of this hypothetical complex. I simulated this process by accelerating Lmod formation in U2017 (Figure 3.9 RVE-OX trace). Interestingly, I observe the expected effect. Any evidence of RVE8 participating in a complex could suggest that the assumption could be correct. Interestingly after proposing this hypothesis it was reported that RVE8 and NIGHT LIGHT-INDUCIBLE AND CLOCK REGULATED (LNK1) and LNK2 form a complex that is important for the activation of RVE8 targets (Xie et al. 2014; Pérez-García et al. 2015). Something remarkable that Pérez-García et al reported is gated association of RVE8 and LNK3,4 proteins at ZT7 by Co-IP in a cross of RVE8-OX/LNK3-OX and RVE-OX/LNK4-OX (Figure 3.10). However, the double mutant lnk1/lnk2 presented higher levels of anthocyanin biosynthesis genes. Which is opposite to the observed down regulation of *PRR5* and *TOC1* in the same mutant (Xing et al. 2015).

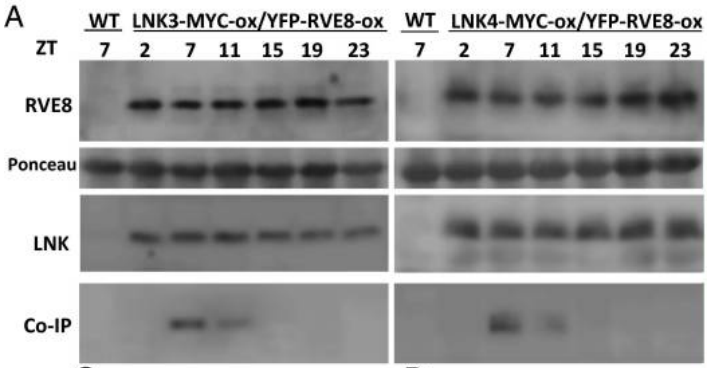


Figure 3.10 Western blot analysis from (Pérez-García et al. 2015) of Co-IP LNK RVE8 with GFP precipitated and detected with anti-MYC from plant overexpressing LNK3-MYC-ox/YFP-RVE8-ox and LNK4-MYC-ox/YFP-RVE8-ox. This allows identification of a complex formation at ZT7 which is dependent on a third component.

This gated association points into the direction that higher levels of RVE8 possible formation of this complex that might act as an activator. Therefore, the RVE8:LNKs:X complex was predicted in P2011, where X is an unknown component or modification to be determined. This theoretical work and reported biological evidence provides a better way for simulating the *lhy/cca1* double mutant in U2017.1. Only the repressive arms of the cL should be eliminated while the activating arms (cLmod) representing the RVEs maintained.

Exploring an alternative mode of regulation for the PRRs

In P2011 the wave of PRRs is a consequence of activation between them, which happens in order CCA1/LHY(cLmod) -> PRR9 -> PRR7 -> PRR5. Experimental evidence now suggests a different situation in which the PRRs

act as repressors (Liu et al. 2016; Liu et al. 2013; Nakamichi et al. 2012; Huang et al. 2012; Gendron et al. 2012; Nakamichi et al. 2010). It is possible to foresee a different mechanism were a wave of inhibition CCA1/LHY(cLmod) l-- PRR9 l-- PRR7 l-- PRR5 l-- TOC1) produces similar results in pealing times to what U2017.1 provides. This was partially introduced in P2012 model were TOC1 inhibited PRR9 (Pokhilko et al. 2013). Other modelling groups have produced more modelling results in this direction (Foo et al. 2016; Fogelmark & Troein 2014). Foo et al showed that the backward repression results in cuspidate (sharp) peaks of gene expression. Therefore, changing the regulation might results in a better fit to the data. I decided to explore this scenario by modifying the regulation of the PRRs so that a backward wave of repression is now responsible for the correct dynamics.

In order to avoid getting lost in the parameter space by introducing this change (remember the mode has more than 104 parameters). I take the following approach. First, I introduce a new set of variables that represent the PRRs. These equations are the revised regulation for the PRRs. The repression and activation from the CCA1/LHY and Lmod (RVE8) variables is maintained in this new equation set, but the equations do not feed-back into the rest of the model. In this way, I avoid modifying the dynamics of U2017.1. The next step consists in fitting these new equations to the PRRs dynamics in U2017.1 (Figuere 3.12).

Once parameters for these new equations are determined the original PRRs transcript equations are swapped by the revised regulation with parameters that reproduce the original PRRs behavior. Then the model is fitted to the original U2017.1 dynamics (Figure 3.13 A). The fitting results show that it is possible to recover U2017.1 profiles, however some discrepancies are observed and surmised in the cost profiling (Figure 3.13 B). We can observe impact in the behavior of PRRs associated variables. This new model receives the name U2017.2. I explain at the end of this chapter how U2017.2 differs from F2014. However, in Figure 3.12 I show how I implemented the regulation of the PRRs.

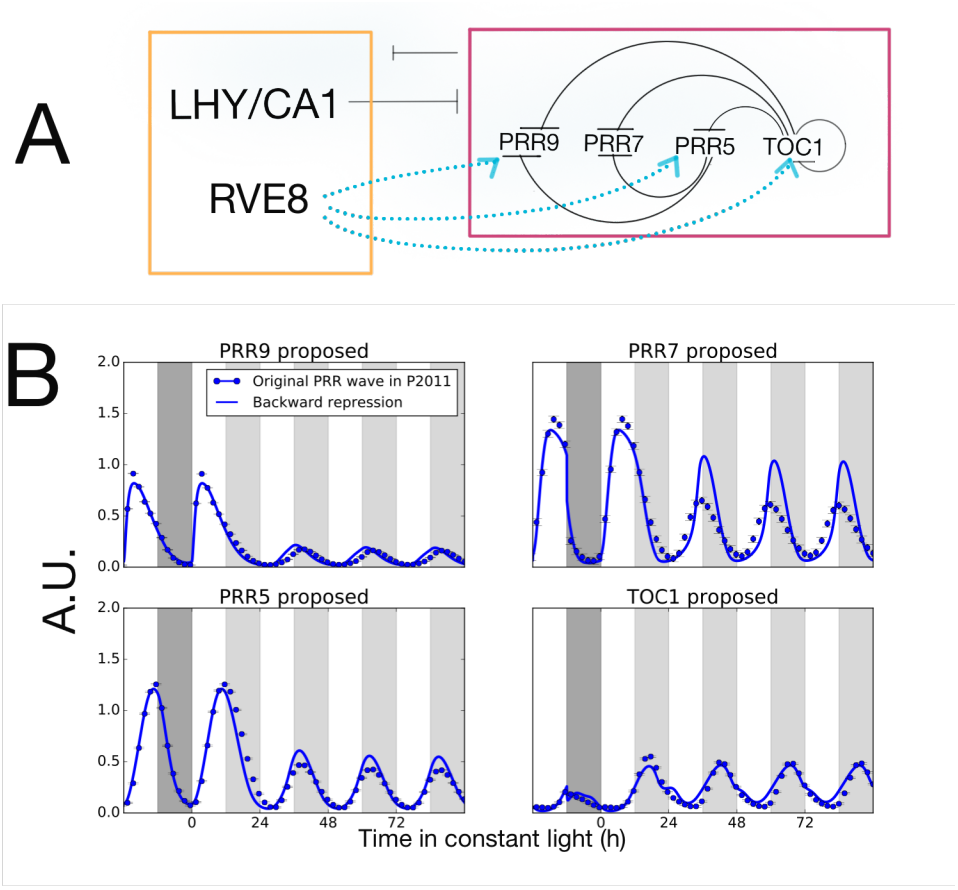
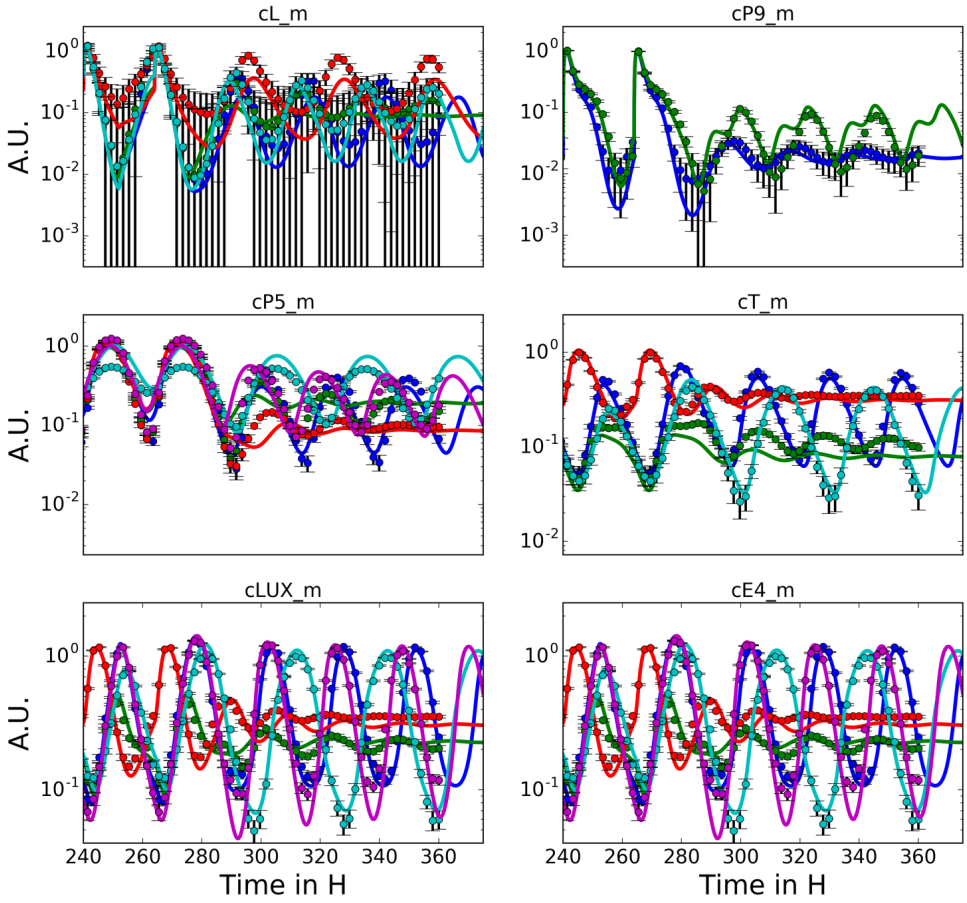


Figure 3.11 Revised PRR regulation and dynamics of proposed equations. A) LHY/CCA1 only acting as repressors, RVE8 activate the expression of PRR9, PRR5 and TOC1. The PRRs now present a backward wave of inhibition. B) Fitting or proposed equations the original PRRs dynamics.



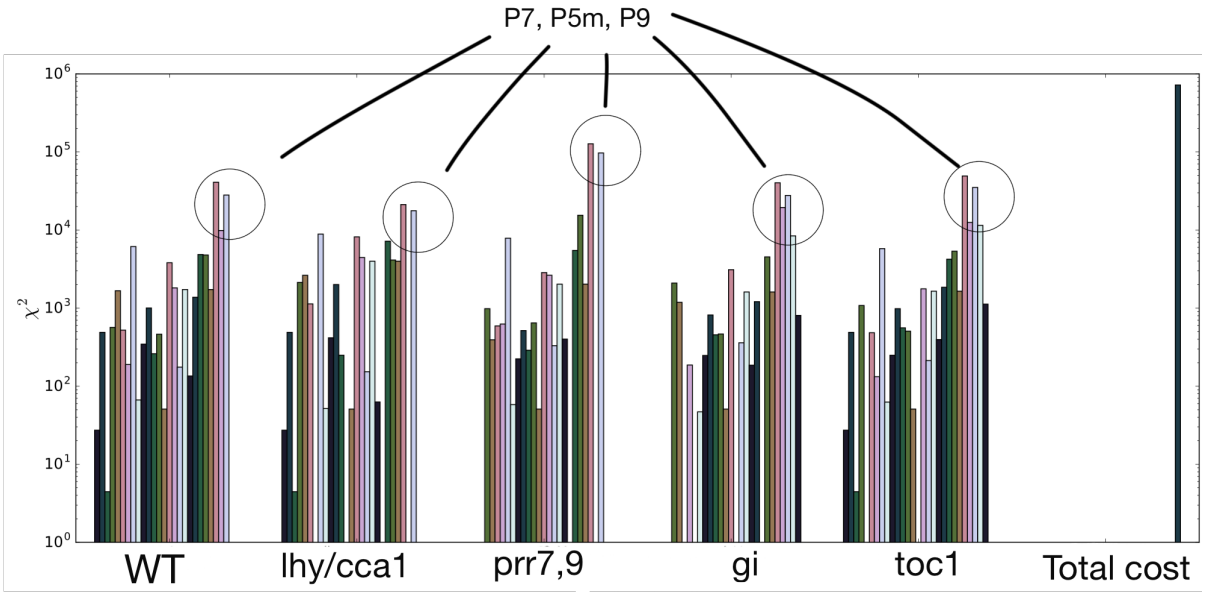


Figure 3.12 Dynamics and cost summary of U2017.2. Upper panels some exemplary genes. Lower panel cost for each time series.

Linking models to qRT-PCR data

After producing the modification on P2011 that resulted in U2017.1 and U2017.2 I linked this model to the original P2011 dynamics by using a likelihood function that assumes normality

$$P(D|M(\theta)) = \prod_{k=1}^{m} \prod_{i=1}^{d_k} \frac{1}{\sigma_{ki}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \frac{(y_{ki} - y_k(t_i, \theta))^2}{\sigma_{ki}^2} \right\}$$

Equation 3.4

Which measures probability of data *D* given model a *M* with parameters set *θ.* Experimental measurements enter as y_{ki} with associated uncertainties σ_{ki}. Model predictions for the measurements y_{ki} are $y_k(t_i,\theta)$. Maximum likelihood estimate for the model parameters can be found by minimising the difference between model and data. This results in Maximum Likelihood estimates for the model parameters.

When linking the U2017 versions, I assumed P2011 dynamics as data and no careful treatment of the uncertainty was performed. Nonetheless this cannot be the case for experimental data. If it does not fulfil the normality assumption, then it will compromise the parameter inference process. The assumption in the inference process is that the model predicts the mean behaviour of smooth data and any deviations are considered experimental noise. I explored if this

assumption holds for Glyceraldehyde 3-ophosphate dehydrogenase GAPDH control from the TiMet data.

Analysis of GAPDH qRT-PCR data reveal a deviation from normality in the experimental error.

Generally, a normality assumption is taken when performing parameter inference using qRT-PCR data. However, no systematic analysis of type of experimental error has been conducted for qRT-PCR using plant material has been reported before. In TiMet quantifications of GAPDH transcript levels were also performed for each time point in the time series, offering a wealth of replicates for assessing what type of distribution the experimental error follows. Furthermore, the levels of the 5' and 3' ends of GAPDH were measured providing more data for robust conclusions. Comparing the qRT-PCR data to the expected theoretical quantiles for a normal distribution shows deviation from normality. These occurs towards the tails of the distribution for two GAPDH regions tested (Figure 3.4 A,B). Natural-log (ln) normal distribution is pervasive in biology. Therefore, I ln-transformed the q-RT-PCR data for both ends of GAPDH and compared it again to the theoretical quantiles of a normal distribution (probability plots) (Figure 3.4 C,D). This transformation normalises the data. A normality test for the ln-transformed data reflects the observation of the probability plots (Table 3.1). This shows that a ln-transformation on qRT-

PCR data generated from *Arabidopsis* tissue for any further analysis, e.g. parameter inference in clock models.

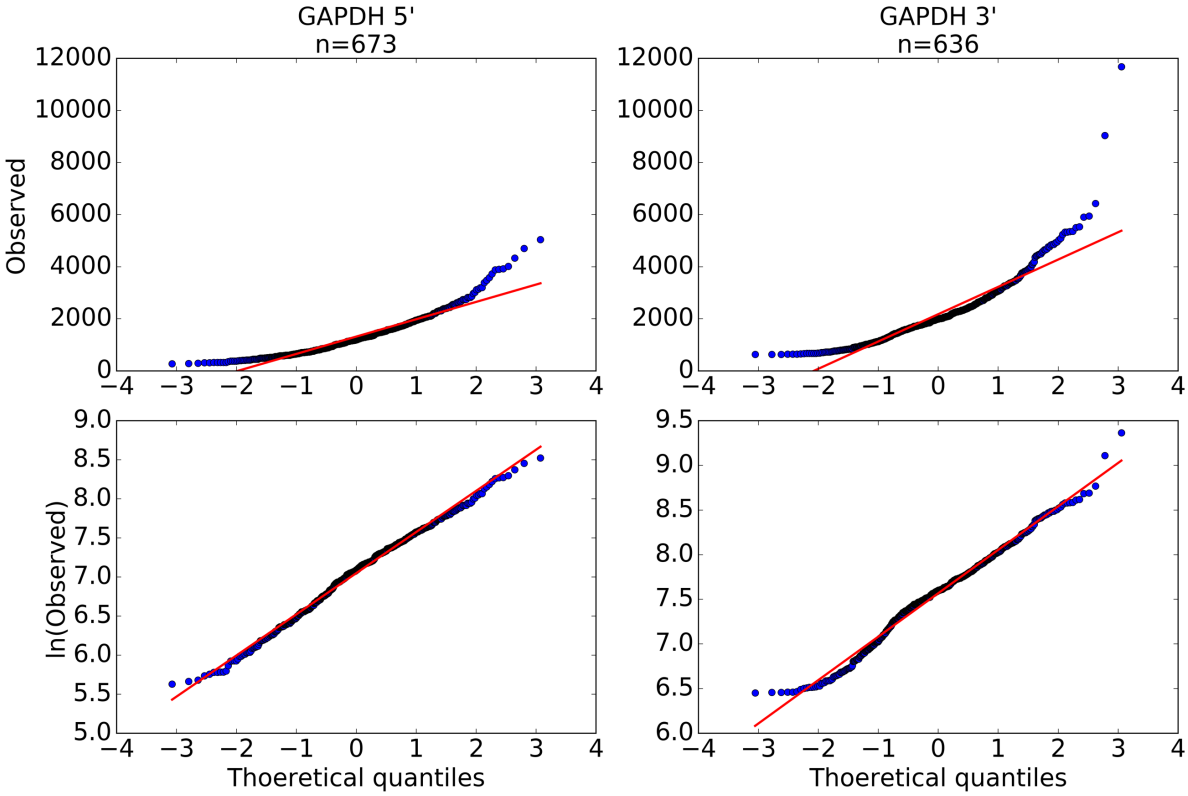


Figure 3.13 Probability plots for GAPDH qRT-PCR data. Upper panels raw data, lower panels natural log transformed qRT-PCR. Theoretical quantiles are derived from the normal distribution. Distribution of GAPDH follow a log-normal distribution. Probability plots for pooled data form TiMet project. Transcript data is better modelled by a log-normal distribution. Data from the Col-0, cca1-1/lhy1-11, prr7/9, toc1 and gi was pooled together.

Table 3.1 Summary of normality test for GAPDH qRT-PCR data. ln denotes data natural log transformed. The test statistic summarizes the z-scores resulting from testing normality in skewness and kurtosis.

Data	**test statistic**	***p-value***
GAPDH 5'	177.09	3.50E-39
ln GAPDH 5'	6.08	0.047
GAPDH 3'	323.83	4.78E-71
ln GAPDH 3'	1.07	0.58

Period constraints

Oscillatory behaviour in constant conditions is a fundamental property of circadian rhythms. In my experience using only the length of the TiMet time-series do not enforce oscillations in constant conditions (not shown). This can be solved by introducing constraints on the oscillation period in constant light condition. Extensive, characterisation of periods in WT and clock mutants provides the data required for enforcing period constraints in constant conditions. Reanalysing public data for two clock marks in WS background CCA1p:LUC and TOC1p:LUC in Biodare2 using FFT-NLS results in a normal distribution for the periods (Figure 3.15). The periods can be approximated by a normal distribution. Therefore, the periods can enter the likelihood (equation 3.5) as additional residuals

Equation 3.5

$$NP(\tau) = \frac{N}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{N}{2} \frac{(\tau - \tau_e)^2}{\sigma^2} \right\}$$

τ is the model period, τ_e is the target period and *N* is a weight introduced to prioritise the period over time series. What I have observed in fitting attempts of all clock models is that a strong weight should be enforced in order to avoid arrhythmic behaviour.

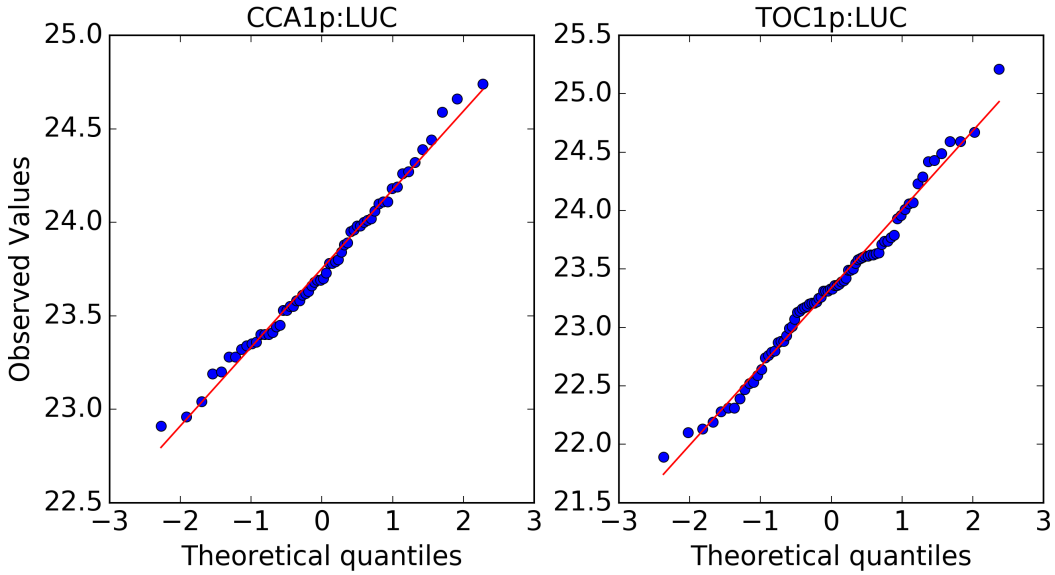


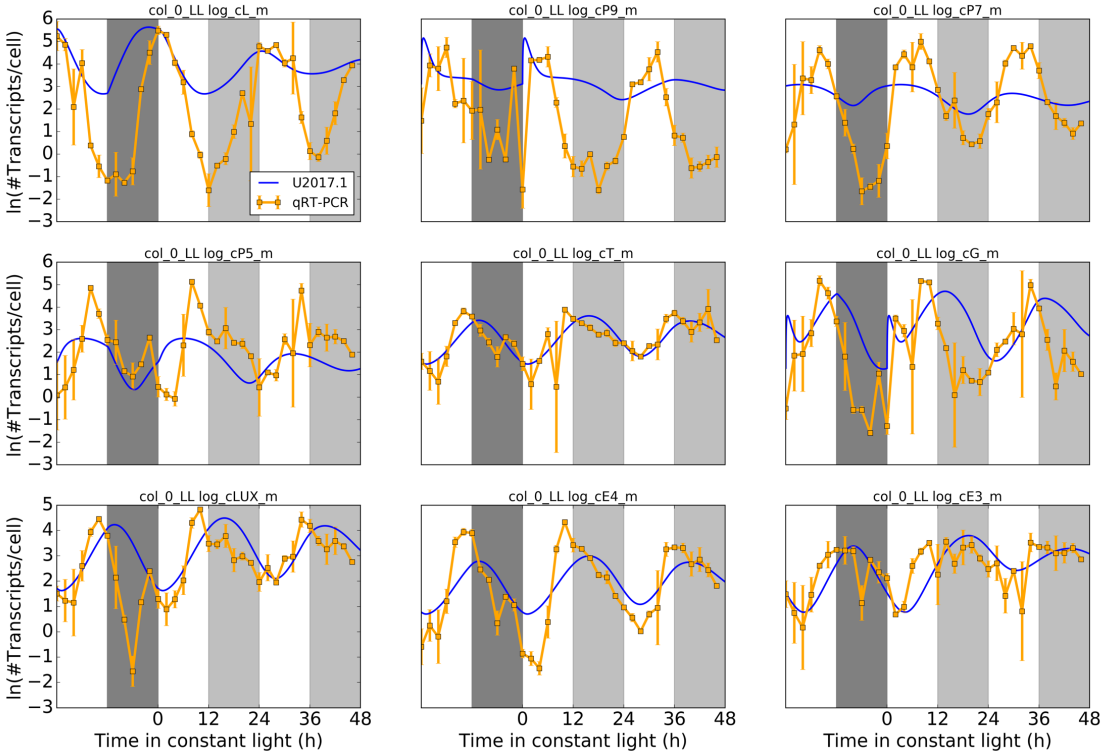
Figure 3.14 Period is normally distributed for two clock markers. Observed period values vs expected quantiles assuming normal distribution for CCA1 and TOC1 promoters fused to LUC. Data analysed in Biodare2 using FFT-NLS method. Biodare experiment id 12730563219125.

Table 3.2 Normalisty test for the period distribution of the circadian observable.

Data	**test statistic**	***p-value***
CCA1:LUC	1.01	0.6
TOC1:LUC	0.79	0.67

Linking models to TiMet data.

I combined the time series and periods to find parameters in absolute units for U2017.1 and .2. The data was ln-transformed and then mean and S.E.M. calculated and used as estimation of uncertainty. I enforced period constraints for each of the mutants after the second day in constant light conditions (Table 3.3). The scaling factors were analytically determined by during the fitting process, therefore only the shape wave shape was used for parameter inference at this stage. Figure 3.16 show a subset of the time-series fitted that correspond to Col-0 data. the global fitting results are presented in (Table 3.4).



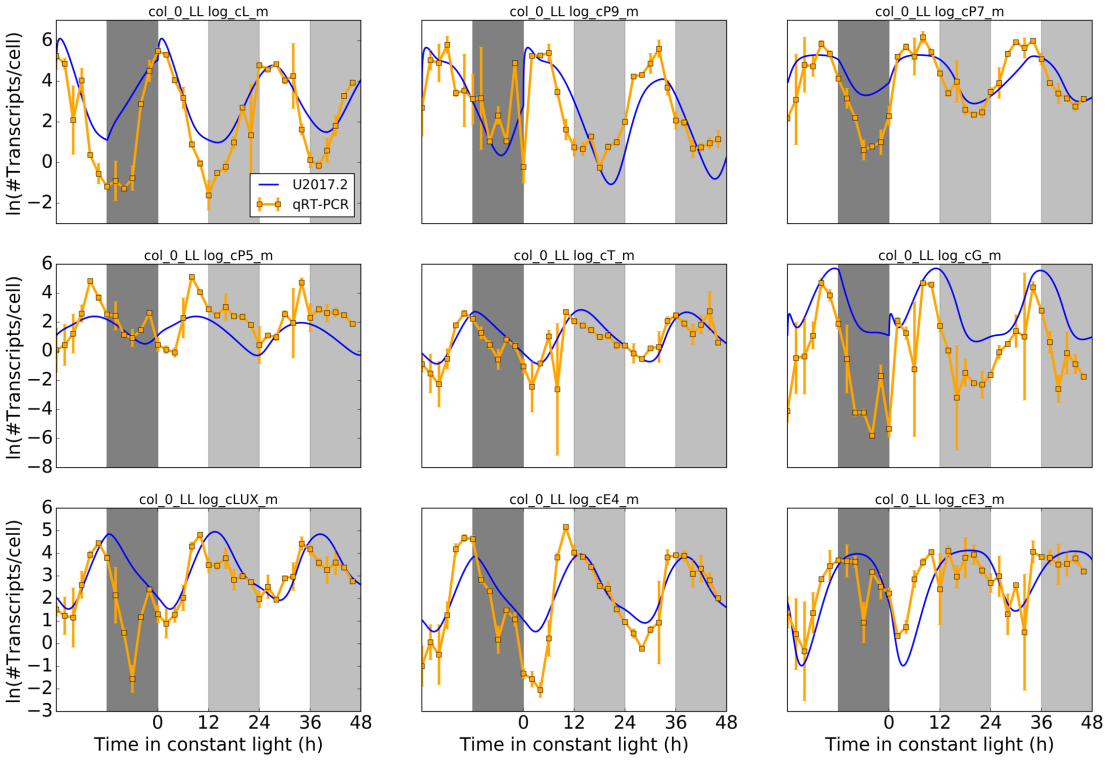


Figure 3.15 Resulting fittings of U2017.1 and 2 to TiMet data. Oscillations were maintained in the WT network with an apparent improvement in the U2017.2 relative to U2017.1. The scale factors were analytically determined for each variable. (Figure 3.17) summarizes the fitting results across all networks.

Table 3.3 Cost of new models to P2011 and TiMet

Model/Cost to	*P2011*	*TiMet*
U2017.1	1.9E4	7.2E6
U2017.2	7.2E5	6.7E6

In this particular example *PRR9* and *PRR7* show better fit in the U2017.2 indicating that the verified PRRs regulation increases the quality of the model. The fitting summary shows that there is general improvement in the model quality relative to U2017.1.

Table 3.4 Periods constraints on for U2017 models and enforced periods

	U2017.1 (τ)	***U2017.2 (τ)***	***Enforced (τ_e)***	***Uncertainty (σ)***
col_0_LL	24.5	24.48	24.5	1E-8
prr79_LL	29.99	29.99	30	1E-8

lhycca1_LL	17.00	16.99	17	1E-9
gi_LL	22.00	22.00	22	1E-8
toc1_LL	20.99	20.98	20	1E-8

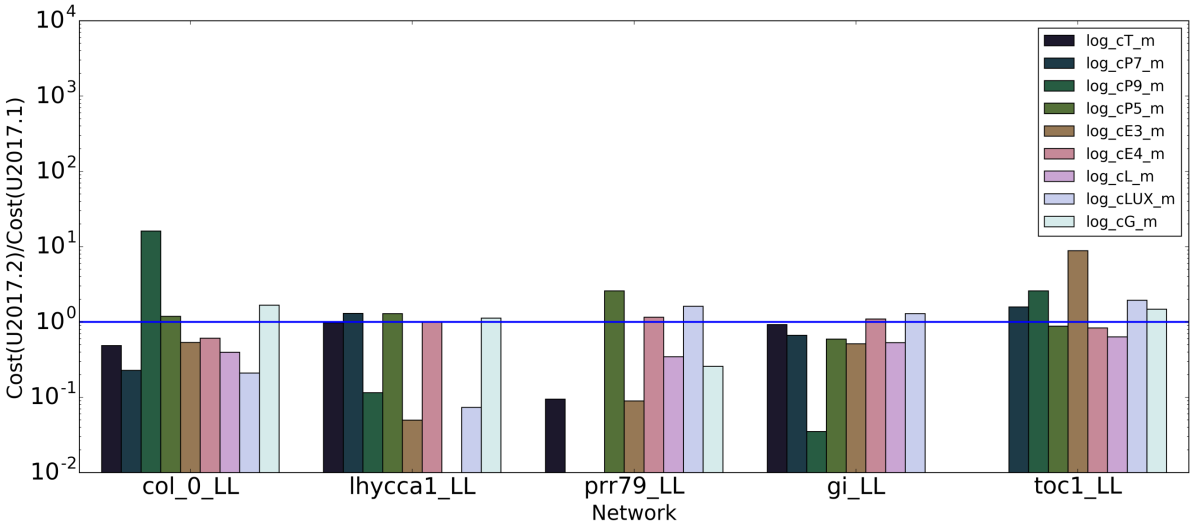


Figure 3.16 Summary of Relative cost of U2017.2 to U2017.1 per variable Summary of fitting to TiMet data for all the networks and time series used. Blue lines means same cost.

Challenging model transcription rates against biological knowledge.

Transcription rates in have never been tested using biologically supported ranges. This reason was the lack of data for performing such analysis. In order to obtain transcription rate ranges that Arabidopsis can produce I integrated two data. The resulting distribution has units of [transcripts]/[cell h].

Transcriptomic measurements of synthesis and decay of RNAs in *Arabidopsis* have been performed RNA labelled with 4sU. That can be purified by thiol-specific biotinylation and strepatavidin-coated magnetic beads. From the incorporation of 4sU into the RNA it is possible to determine the rate of

transcript synthesis, degradation and transcript steady state levels (Figure 3.18) (Sidaway-Lee et al. 2014). Nonetheless, microarrays were used for determining RNA levels in these experiments, results in data with units [microarray units]/[h]. In order to transform this rates into [transcripts]/[cell h], I explored the use of qRT-PCR measurements calibrated for [transcripts]/[cell] (Piques et al. 2009).

The qRT-PCR data consisted of 96 transcripts of central metabolisms measured at two time points, denominated light faction and dark fraction (Figure 3.19 A). First, I extracted transcripts that show less than 0.2-fold change between light and dark fraction in the Piques set. Then, I assume that these transcripts are in steady state. This operation provides a linking path for the steady state levels of Sidaway-Lee microarray data (Figure 3.19) to the steady state levels of Piques qRT-PCR measurements (Figure 3.20B).

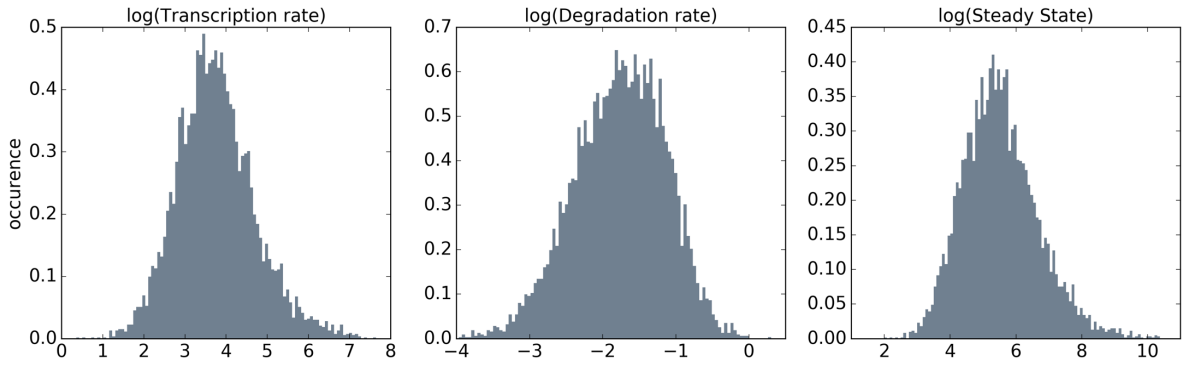


Figure 3.17 Steady state levels, transcription rates and decay rates determined by Sidway-Lee et al 2014. All the distribrions were log transformed. Steady stated (C) was dtermined by dividing synthesis rates (A) by degradation rates (D).

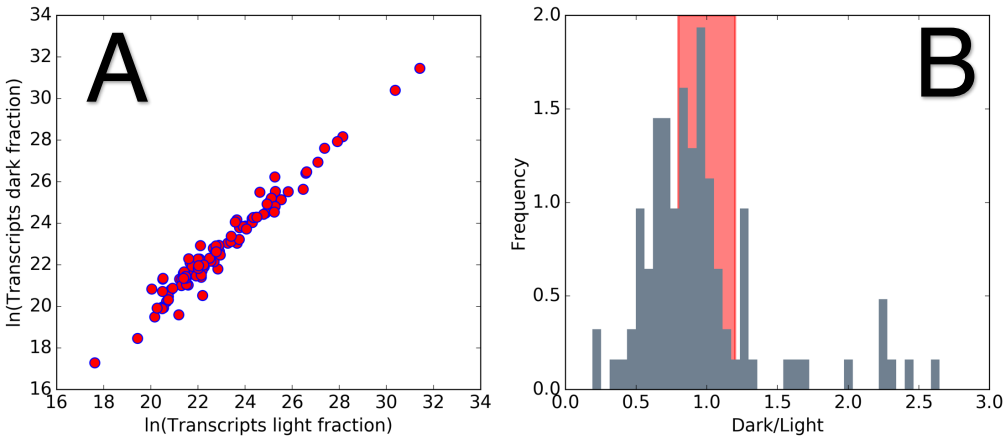


Figure 3.18 Gene selection for calibrating microarray data. A) Transcript level in light and dark large polysomal fraction in Piques dataset. B) Red band are the genes selected with less than 0.2 fold change between light and dark conditions

Assuming no regulation on the transcription for all the genes in the microarray takes place then we can write the following dynamic model

$$\frac{dS_m}{dt} = a_s - bS_m$$

Equation 3.6

Where S_m stands for the transcript levels in Sidway-Lee data at steady sate, a_s stands for the transcription rate and b for the transcript decay constant in Sideway-Lee. At steady state (setting the eq 3.5 to 0) and solving for a_s

$$a_s = bS_m$$

Equation 7

When considering equation 3.5 in the Sidaway-Lee context a_s has units [microarray units]/[hour], In Piques context a_p has units [transcripts]/[h cell. Linking between the two contexts can be done by performing a linear regression Sidaway-Lee vs Piques (< 0.2 fold change transcripts). The resulting regressing model is substituted in the Piques quasi-steady state equation resulting in

$$a_p = bP_m$$
$$a_p = b(mS_m + i)$$

Equation 8

a_p is transcription rate with units in [transcripts]/[h cell]. In principle, the decay constant *b* should be identical between the two data sets with units 1/[h], and it is specific for each transcript. the slope *m* has units [transcripts]/[microarray units] and the intercept *i* has units of [transcripts]. Remember from last section, that the regression has to be done with ln-transformed data. Therefore, the correct form of equation 7 is,

$$a_p = b(exp(m\ln(S_m) + i))$$

Equation 9

Sidaway-Lee reported microarray data for two different temperatures, the linear regression with better fit is the one close to the experimental conditions

in which Piques data was collected which is an encouraging observation (Figure 3.20).

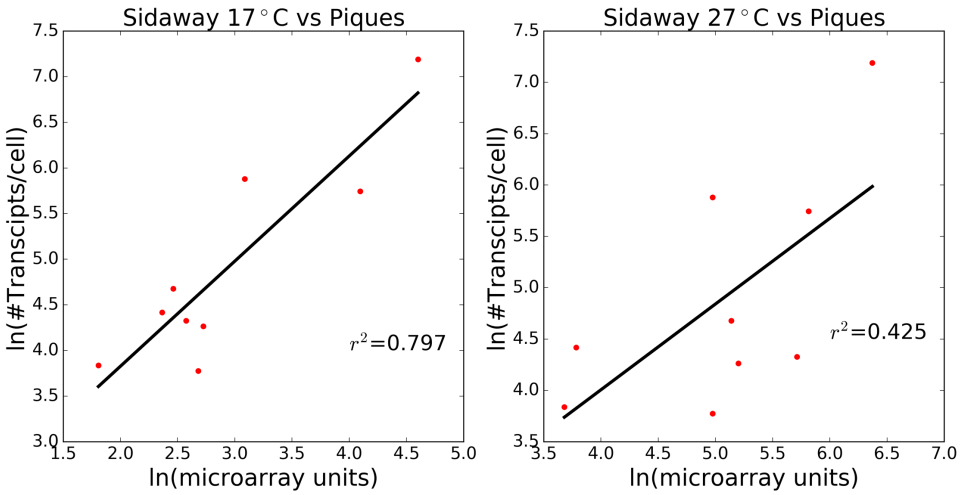


Figure 3. 19 Calibration of Sidaway lee data with qRT-PCR data in absolute units. A) Regression with transcription rates at 21ºC using genes with 0.8 fold change, B) Regression of transcription rates at 27ºC using genes with < 0.2 fold change. More supported correlation for 21ºC compared to 27ºC

After performing the regression, I calibrate each of the measurements present in the microarray resulting in a distribution with units in [transcripts]/[cell h] (Figure 3.21)

Having this distribution is a major step for constraining clock models. For the first time, it is possible to evaluate the biological soundness of the inferred rates. Therefore, I rescaled the U2017.1 and U2017.2 to match the TiMet data scale in addition to the wave form. This distribution can be used for

constraining transcription rates during parameter inference or testing the current parameters in plant mathematical models.

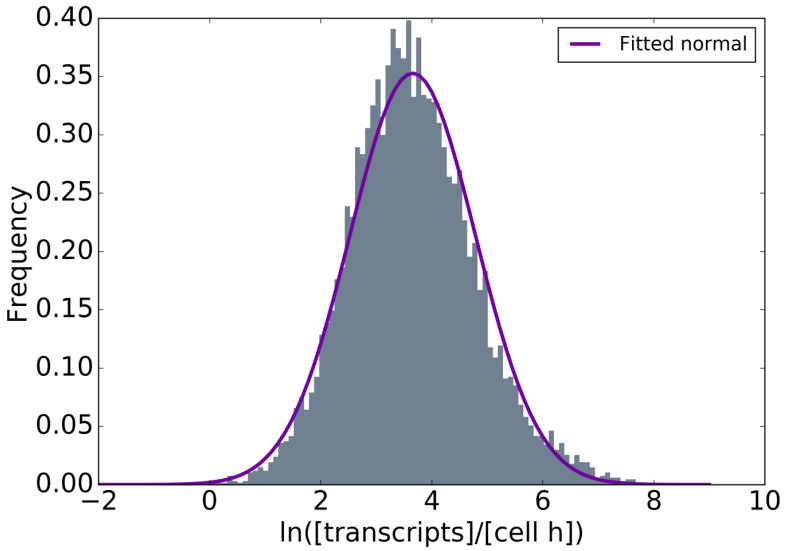


Figure 3.20 Calibrated Siday-Lee microarray data. Distribution of transcription rates across the genome in at 17ºC with units [transcripts]/[cell h].

Then, I tested U2017.1 and .2 by rescaling the fitted parameters for matching the mass units of TiMet. The results are models with transcription rates [Transcript]/[h cell]. Because I already fitted the models to the shape of TiMet data I decided to avoid modifying the dynamics by introducing scaling factors in the equations. This allow me to exploit the symmetries in the model for doing a change in units without modifying the waveform. I introduce a new variable

$$scT_m = cT'_m$$

Equation 10

Where *s* is the scaling factor that needs to be propagated into the model by substituting the original variable by the rescaled one *cT_m'*. Taking the derivative for this new variable with respect to time results in,

$$s\frac{dcT_m}{dt} = \frac{dcT'_m}{dt}$$

Equation 11

Then, substituting the variable and its derivative into the transcription rates of the models results in rescaled equations,

$$\frac{1}{s}\frac{dcT'_m}{dt} = nf(cN_1, cN_2, \ldots, cN_n) - \frac{m}{s}cT'_m$$
$$\frac{dcT'_m}{dt} = snf(cN_1, cN_2, \ldots, cN_n) - mcT'_m$$

Equation 12

Therefore, the transcription rate *n* now its scaled by *s*. The transcript variables are also present in the protein synthesis rates. This part of the model has to be also updated with the new variable *cTm'*. The result is an adjusted rate of translation which balances the effect of a different amount transcript at the translational level.

$$\frac{dcT}{dt} = \frac{p}{s} cT'_m - mcT$$

Equation 13

The translation rate *p* is now modified by the *s*. *cT* concentration of transcription factor *T*, *m* decay constant of *T*. The distribution of rescaled transcription rates for U2017.1 and U201.2 are presented in (Figure 3.23). The ln-normality allows calculating z-scores for model rates (Table 3.5). Interestingly, even though the U2017.2 produces better fits than U2017.1 the transcription rates required by .2 are more extreme than for .1, resulting in a more unrealistic picture. Also, most of the changes in transcription rate happen for the PRR elements. The evening complex elements retain similar values between the two models. U2017.1. The better fitting in the revised PRR version comes with a cost in transcription rates.

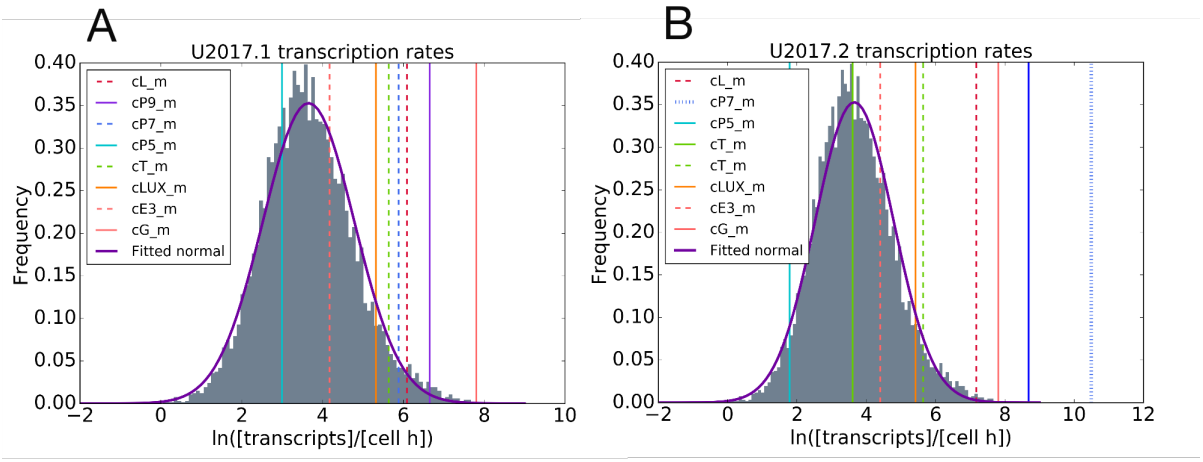


Figure 3.21 Distribution of rescaled transcription rates that match TiMet data in A) U2017.1 and B) U2017.2 compared to the integrated distribution

(Urquiza) of Piques and Sideway. The model rates were ln-transformed and plotted.

Table 3.4 z-scores for the transcription rates in the tested models for each gene.

	U2017.1	***U2017.2***
cE4_m	1.47	1.56
cE3_m	0.46	0.66
cP5_m	-0.58	-1.65
cP7_m	6.04	6.04
cT_m	1.76	-0.04
cP9_m	2.65	4.45
cG_m	-0.13	3.67
cL_m	2.15	3.11
cLUX_m	1.47	1.56

Discussion

In this chapter I explored for the first time the use of absolutes units in *Arabidopsis* clock models at the transcriptional levels for the first time. I introduced two new models U2017.1 and U2017.2. The first one removes the quasi-state assumption for intermediaries of the EC. I provided some new luciferase data regarding BOA that suggest it could potentially help in extending the activity of the evening complex to the end of night. I proposed a reinterpretation of the cLmod variable, mapping it to RVE8 transcriptional activator. The modelling results presented for Lmod suggest that RVEs were already predicted in P2011. The reinterpretation of cLmod then using U2017.1 supports the hypothesis that a complex formation takes place at ZT6 (Pérez-García et al. 2015).

A concern for P2011 is the regulation of the PRRs in which the wave of expression is generated by activation rather than repression. A revised version of for TOC1 was already explored in previous in the lab. However, F2014 and MF2016 explored this addition and its impact in the dynamics of the clock. Nonetheless, no intermediate states in the full development of their models exist. In the post-P2011 models the introduction of new elements and molecular species complicates studying the impact of use a revised regulation for the PRR transcriptional regulation. The original PRR regulation in P2011 has been criticised as new experimental evidence suggest that these proteins act as repressors. Implementing this regulation using as base model P2011 resulted in U2017.2. Figure 3.23 shows the differences between F2014 and U2018. The latter network presents less complexity in several aspects. Only total pools of PRRs are represented, implying that the PRRs act immediately in the nucleus as they are produced. LHY/CCA1/RVE8 are clustered in a single variable. No autoregulation of LHY has been considered. The model ignores the presence of the BROTHER OF LUX ARRYTHOMO (BOA) also known as NOX. RVE8 activation only takes place on PRR9, PRR5 and TOC1. The evening complex is only regulated by CCA1/LHY. In addition to a decrease in model complexity, I decided to maintain the translation rate as explicit parameters. This will be used in further work when introducing absolute numbers at the protein level.

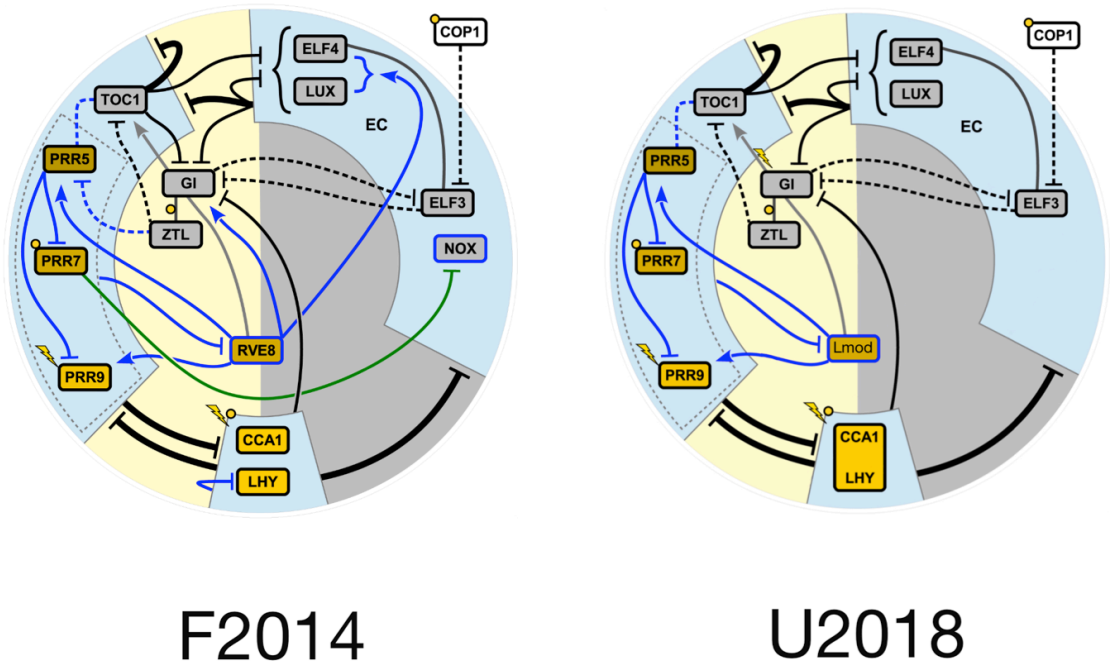


Figure 3.22. An intermediate model between P2011 and F2014. I generated U2017.2 In order to explore the impact of modifying how PRRs type of regulation impact model performance. The model has less links than F2014 however it retains translation rates as free parameters. The PRRs work as repressors that repress each other with TOC1 acting on all the PRRs.

Fitting U2017.2 to the TiMet data shows that modification of PRR regulation result in better fittings. However, the resulting transcription rates after this fitting fall become unrealistic. Interestingly these unrealistic rates map to the PRRs. The transcription rates for the E.C. components are similar between the two tested models. This conservation of parameter values might reflect the fact the PRRs do not directly regulate *LUX*, *ELF3* or *ELF4.*

In mice Jolley et al used for the first-time empirical distributions of parameters for constraining the fitting process (Jolley et al. 2014). In a similar way analyzing data from Sidaway Lee and Piques I was able to derive a distribution of transcription rates that can be used for constraining transcription rates in future computational work. The distribution can be used in the context of the clock or when modelling other biological process in *Arabidopsis.* In this chapter I used as model system P2011 and small modified versions however introduction of absolute units in F2014 might be ideal in future computational work. However, the work presented in this chapter provides the roadmap for achieving that in more complex models at the transcriptional level.

References

Chew, Y.H. et al., 2017. Linking circadian time to growth rate quantitatively via carbon metabolism. *bioRxiv*, p.105437.

Dai, S. et al., 2011. BROTHER OF LUX ARRHYTHMO is a component of the Arabidopsis circadian clock. *The Plant cell*, 23(3), pp.961–972.

De Caluwé, J. et al., 2016. A Compact Model for the Complex Plant Circadian Clock. *Frontiers in plant science*, 7, p.74.

Domijan, M. & Rand, D.A., 2015. Using constraints and their value for optimization of large ODE systems. *Journal of the Royal Society, Interface / the Royal Society*, 12(104), pp.20141303–20141303.

Domijan, M. et al., 2016. PeTTSy: a computational tool for perturbation analysis of complex systems biology models. *BMC bioinformatics*, 17(1), p.124.

Flis, A. et al., 2015. Defining the robust behaviour of the plant clock gene circuit with absolute RNA timeseries and open infrastructure. *Open biology*, 5(10), p.150042.

Fogelmark, K. & Troein, C., 2014. Rethinking transcriptional activation in the Arabidopsis circadian clock. *PLoS computational biology*, 10(7), p.e1003705.

Foo, M., Somers, D.E. & Kim, P.-J., 2016. Kernel Architecture of the Genetic Circuitry of the Arabidopsis Circadian System. L. J. Jensen, ed. *PLoS computational biology*, 12(2), p.e1004748.

Gendron, J. et al., 2012. Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8), pp.3167–3172.

Hayama, R. et al., 2017. PSEUDO RESPONSE REGULATORs stabilize CONSTANS protein to promote flowering in response to day length. *The EMBO journal*, 36(7), pp.904–918.

Helfer, A. et al., 2011. LUX ARRHYTHMO encodes a nighttime repressor of circadian gene expression in the Arabidopsis core clock. *Current biology : CB*, 21(2), pp.126–133.

Hsu, P. & Harmer, S., 2012. Circadian phase has profound effects on differential expression analysis. *PloS one*, 7(11).

Huang, W. et al., 2012. Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator. *Science (New York, N.Y.)*, 336(6077), pp.75–79.

Jolley, C.C. et al., 2014. A mammalian circadian clock model incorporating daytime expression elements. *Biophysical journal*, 107(6), pp.1462–1473.

Keily, J. et al., 2013. Model selection reveals control of cold signalling by evening-phased components of the plant circadian clock. *The Plant Journal*, 76(2), pp.247–257.

Liu, T. et al., 2013. Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. *The Plant Journal*, 76(1), pp.101–114.

Liu, T.L. et al., 2016. A G-Box-Like Motif Is Necessary for Transcriptional Regulation by Circadian Pseudo-Response Regulators in Arabidopsis. *Plant physiology*, 170(1), pp.528–539.

Nakamichi, N. et al., 2010. PSEUDO-RESPONSE REGULATORS 9, 7, and 5 are transcriptional repressors in the Arabidopsis circadian clock. *The Plant cell*, 22(3), pp.594–605.

Nakamichi, N. et al., 2012. Transcriptional repressor PRR5 directly regulates clock-output pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42), pp.17123–17128.

Nusinow, D. et al., 2011. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

Pérez-García, P. et al., 2015. Time-dependent sequestration of RVE8 by LNK proteins shapes the diurnal oscillation of anthocyanin biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 112(16), pp.5249–5253.

Piques, M. et al., 2009. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. *Molecular systems biology*, 5, p.314.

Pokhilko, A. et al., 2012. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular systems biology*, 8, p.574.

Pokhilko, A., Mas, P. & Millar, A., 2013. Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. *BMC systems biology*, 7, p.23.

Rawat, R. et al., 2011. REVEILLE8 and PSEUDO-REPONSE REGULATOR5 form a negative feedback loop within the Arabidopsis circadian clock. G. P. Copenhaver, ed. *PLoS genetics*, 7(3), p.e1001350.

Schmal, C., Reimann, P. & Staiger, D., 2013. A Circadian Clock-Regulated Toggle Switch Explains AtGRP7 and AtGRP8 Oscillations in Arabidopsis thaliana. *PLoS computational biology*, 9(3).

Seaton, D.D. et al., 2015. Linked circadian outputs control elongation growth and flowering in response to photoperiod and temperature. *Molecular systems biology*, 11(1), pp.776–776.

Sidaway-Lee, K. et al., 2014. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. *Genome biology*, 15(3), p.R45.

Xie, Q. et al., 2014. LNK1 and LNK2 are transcriptional coactivators in the Arabidopsis circadian oscillator. *The Plant cell*, 26(7), pp.2843–2857.

Xing, H. et al., 2015. LNK1 and LNK2 recruitment to the evening element require morning expressed circadian related MYB-like transcription factors. *Plant signaling \& behavior*, 10(3), p.e1010888.

Chapter 4
Theoretical Introduction of Absolute Units for Protein Variables in the *Arabidopsis* Circadian Oscillator

"I demonstrate by means of philosophy that the earth is round, and is inhabited on all sides; that it is insignificantly small, and is borne through the stars" –J. Kepler-

The use semi-quantitative or fully quantitative protein time series in mammalian and fungal clock models have helped to gain mechanistic insight. Quantitative insight has challenged simplifying assumptions and equivalence of clock architectures between organisms (Lee et al. 2001; Merrow et al. 1997; Forger & Peskin 2003; Kim & Forger 2012). Further development of quantitative methodologies for clock proteins resulted in absolute quantifications for the animal clock (Narumi et al. 2016). However, this latest data has not been incorporated in mathematical models for the mammalian system.

Unfortunately, in *Arabidopsis* protein data for clock genes is scarce and generally not suitable for performing detail quantitative studies. The only reported detail absolute quantifications that could be used in clock modelling

is for the Phytochrome photoreceptors (Sharrock & Clack 2002). The technical challenge associated with the generation of quantitative protein data to some extent justify the use of RNA as the most direct readout of the internal state of the oscillator. One of the most recent improvement in RNA data was the absolute quantifications used in Chapter 3. Therefore, past modelling work had to deal with the lack of high-quality data that measures the abundance of the Arabidopsis clock proteins over daily cycles, in a semi quantitative or fully quantitative manner.

As stated in early in this thesis I am interested in using absolute units for uncovering potentially unrealistic assumptions in clock models by incorporating data in absolute units. However, before any experimental results can be incorporated, several, linked developments were required in the model and in theoretical methods, in order to prepare the current clock models to use such data. Including methods for judging the quality of experimental quantification, similarly to (Simicevic et al. 2013).

The focus is whether U2017.2 protein parameters are biologically reasonable values? I start addressing this problem by hypothesising protein levels by using a simple data driven model of translation combined with absolute quantification of clock transcripts (Chapter 3) (Flis et al. 2015; Flis et al. 2016). Using the predicted protein levels, I rescaled the protein component of U2017.

This rescaling is more challenging than for transcripts, mainly because of the protein interactions the model presents.

The proteins in U2017 are produced in a simple fashion from their cognate transcripts. But more heterogeneous processes are modelled at the post-translational level. For example, PRR protein degradation can be controlled by light or other components e.g. ZIETELUPE (ZTL). The Evening Complex is formed by ELF3, ELF4 and LUX and the flux of this proteins into the complex formation is modulated by other clock components (GIGANTEA, GI) or the photomorphogenic regulator CONSTITUTIVE PHOTOMORPHOGENESIS (COP1) in combination with light. This complexity requires a subtler treatment in the rescaling.

After performing the rescaling, the question that follows is: How can I judge the validity of the predicted protein levels? An important assumption in the mechanistic model is that many of the elements are transcription factors and their action at promoters is assumed to be in equilibrium. I use these equilibrium assumptions to link model parameters to data and to judge the predicted protein levels. I consider several data types to do so, including affinity data (Surface Plasmon Resonance or Electrophoretic Mobility Shift Assay), Protein Binding Microarrays, Chip-seq experiments and volumetric measurements. In particular, I unify all these data elements through a biophysical model for assigning affinities to promoter regions using some of

the ideas from (Kinney et al. 2007). In essence in this chapter I prepare U2017.2 for incorporating protein data in absolute units. Which will also help in judging the quality of protein quantification for Arabidopsis clock proteins.

Results

Predicting protein levels from absolute transcript quantification

In order to circumvent the issue of protein data without absolute units I decided to use a simple model (SM) driven by RNA data, to produce a starting estimate of protein levels, which could be used to re-scale protein components in the U2017 model. The general form of the SM is

$$\frac{dP}{dt} = nM(t_i) - mP$$

Protein concentration *P*; *n* translation rate of interpolated transcript data; $M(t_i)$, the amount of interpolated transcript at time t_i and protein decay constant *m.* The transcript data is the absolute quantification of clock transcripts presented in Chapter 2. The task that remains is estimating the translation and decay constants, *n* and *m*. In the case that evidence for light regulation exists, the decay expression used is $m_1+m_2(1-light)$. In that case $m_2=m_1-m_2^{exp}$, m_1 is the degradation measured in light and m_2^{exp} the dark measured decay. Estimated decay constants have been determined for several clock proteins, (CCA1, PRR9, PRR7 and PRR5) (Song & Carré 2005; Kiba et al. 2007; Ito et al.

2007). In the case of TOC1 the decay constant was not quantified but a western blot has been reported I used imageJ64 to obtain an approximation to the decay constant from this western blot, this inferred value incorporates the action of (ZEITELUPE) ZTL protein (Más et al. 2003).

No decay constants have been determined for LUX, ELF3 or ELF4 proteins. However, figures of western blot time-series were available from the literature. Then, I approached the issue by fitting the predicted protein levels to the western profiles by minimising the cost function

$$\chi^2 = \frac{1}{2}\sum_{i=1}^{N_r}(By(t_i,\theta) - Y_i)^2$$

Equation 4.1

Where $y(t_i,\theta)$ is the model prediction at time t_i. Y_i is the data value at t_i. B is the scale factor and N_r the total number of western blot data points. The scaling factor B can be determined analytically by taking the derivative of the cost function with respect to B, setting it to 0 and then solving for B (Brown & Sethna 2003), which results in

$$B = \frac{\sum_{i=1}^{N_r}(y(t_i,\theta)Y_i)}{\sum_{i=1}^{N_r} y(t_i,\theta)^2}$$

Equation 4. 2

This shows that I only need to fit the decay constants, for LUX, ELF3 and ELF4. The translation rates are not fitted as they introduce the units of mass. I assume a very simple translation rate that is proportional to the amount of RNA. I am aware that several factors can affect these rates. For example light affects polysome loading, resulting in higher translation rate in light compared to darkness (Piques et al. 2009). Nonetheless, I will ignore all these factors as a first approximation in the SM. The model could be refined in future work if required.

In this work translation rates will depend on the number of codons (CDS dependent), the ribosomal density and ribosome elongation speed. These values are held constant across the time-series. The ribosome elongation speed was set to 3 codons/s similar to (Piques et al. 2009), which used as reference the elongation rate of animal cells of 4-5 amino acids/s at 25-26ºC. They corrected this rate by using the temperature coefficient Q_{10}, which quantifies the rate of change of biochemical reactions when the temperature is varied by a factor of 10ºC,

$$Q_{10} = \left(\frac{R_2}{R_1}\right)^{\frac{10^\circ C}{T_2 - T_1}}$$

Equation 4. 3

Where R_2 and R_1 are the rate of two reactions and T_2 and T_1 are the temperatures at which R_2 and R_1 are happening. Inverting this function and solving for the reaction R_1 occurring at the lower temperature R_1 results in,

$$R_1 = R_2 Q_{10}^{\frac{T_1 - T_2}{10^\circ C}}$$

Equation 4 4

In particular I used the mean values 4.5 a.a./s and T_2=25.5ºC, T_1=21ºC and a Q_{10} of 2.5. As I mentioned earlier I do not assume differential polysomal loading between light and darkness. Most of the RNA time points are in LL, therefore I decided to use the same assumption as piques for the light regime, that the transcripts will be in the large polysomal fraction, hence assuming 6 ribosomes per kb transcript. The CDS of each of the genes was obtained from TAIR10. These parameters were combined to estimate the rate of translation.

The numerical procedure for using the RNA data is also relevant. The data consists of discrete transcript measurements that if fed directly into the numerical solver would result in lower estimates of protein numbers. Therefore, I log-transformed the data and performed a cubic interpolation to

produce a continuous RNA timeseries trace. Log-transforming the data avoids the issue of negative values when performing the interpolation. Figure 4.1 shows as an example the predictions for CCA1 and LHY proteins that will be used for rescaling U2017. The procedure is similar for the other genes for which transcript data is available.

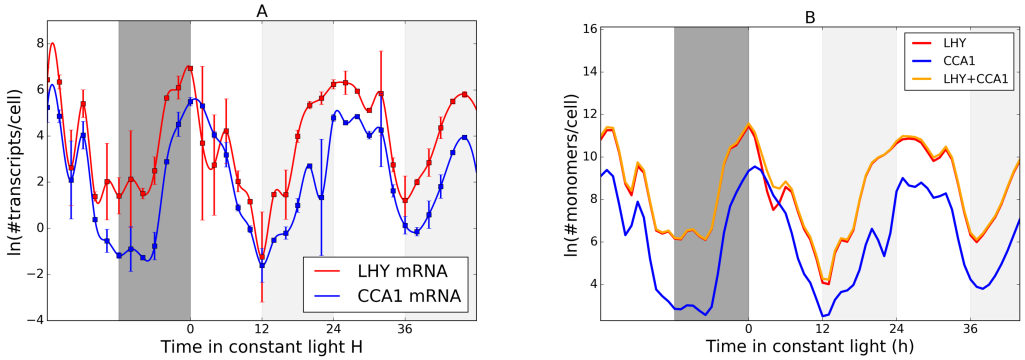


Figure 4.1 Construction of synthetic data for LHY and CCA1 proteins. Cubic interpolation of ln-transformed RNA data of Flis et al, 2015 (A) provide the input for a simple translation model, predicting ln-transformed levels of CCA1(blue) and LHY (red) proteins, and their sum (LHY + CCA1)(orange) (B) in diurnal conditions followed by constant light. Grey background, darkness; white background, light CCA1 $t_{1/2}$= 0.2, LHY $t_{1/2}$=0.39.

Rescaling *cL* to match the predicted levels of LHY and CCA1 protein

In U2017 LHY and CCA1 proteins are lumped together in a single variable *cL*. But these genes present approximately a three-fold difference in transcript level at ZT2. The protein prediction for each of them presents an 8-fold difference at that time (Figure 4.1). These genes were originally lumped as they are thought to present some level of redundancy. For example, neither single mutant shortened period as severely as the double mutant. CCA1 and

LHY proteins form homo and heterodimers; SPR data gave an that their affinity for consensus promoter sites was ~1.14 nM (Lu et al. 2009; O'Neill et al. 2011). In U2017 the way *cL* acts already considers the dimer formation, because the Hill coefficients for promoter binding are set to 2. Then the model simulates the total amount of CCA1 and LHY protein as if they form the same protein pool. I therefore performed the rescaling using the sum of predicted CCA1 and LHY.

Similar to what is done for the RNA rescaling in Chapter 3, I introduce a new variable *cL'*

$$cL' = \frac{1}{\gamma} cL$$

Equation 4. 5

Maintaining the dynamics in this case requires adjusting the dissociation parameters of *cL* for its target promoters, which is the natural result from the introduction of *cL'.* The starting model in this chapter is U2017.1. After rescaling the protein component, I will refer to the model with new parameters as U2017.1.2. I numerically obtained the U2017.1.2 scaling factors to match the protein levels predicted by the simple model using equation 4.1, but setting B=1, so that the U2017.1 model parameters must change in order to produce desired protein levels without a further rescaling factor. This results in *cL* levels around 6×10^4 (Figure 4.2A). But I notice that the synthetic protein data has

higher levels than the rescaled U2017. This is likely due to a temporal mis-alignment of model to data, which results lower levels of cL giving the best cost for U2017. As level is more important than detailed timing in this case, I changed the approach by matching the levels at peak time between the synthetic data and *cL* (Figure 4.2B).

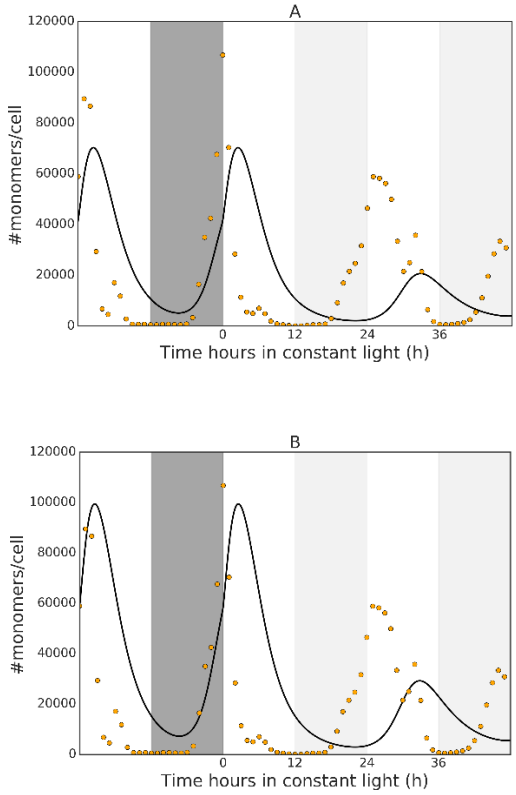


Figure 4.2 Matching U2017 cL to synthetic data. A) U2017 (solid black) misalignment of cLm and cL results in poor rescaling to synthetic data (yellow circles). B) Use of maximum level as target rather than whole synthetic data time-series.

It is important to note that the degradation of *cL* is more complex in U2017 than in Simple Model (SM). In particular cL is transformed into cLmod (RVE8). This "subtle" difference complicates comparing the parameters the

degradation. Instead I analyzed the ln of the ratio of synthesis and degradation in both models (Figure 4.3). The results show that U2017.1.2 synthesis and degradation are less extreme compared to Simple Model for generating the same protein levels. This result, I suggest that the rates for cL in U2017.1.2 are might be within biological ranges.

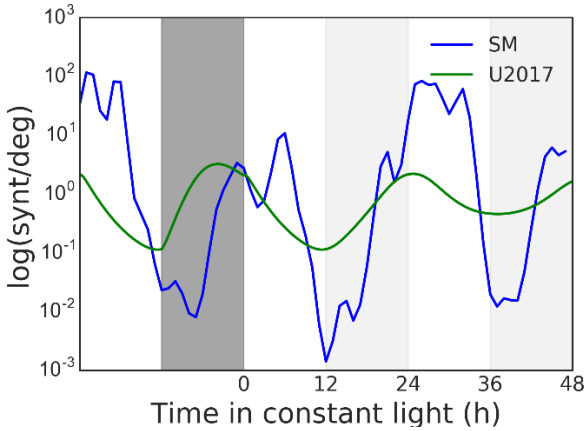


Figure 4.3 Ratio of synthesis and degradation in U2017.1.2 and the Simple Model (SM). The rate of synthesis in SM was calculated by using the interpolated RNA levels. The rate of proteins synthesis for U2017.1.2 by simulating the protein levels and substituting the results back into the model equations.

The impact of rescaling on the dissociation constants for the promoters of cL targets

As mentioned in the introduction, I am interested in assessing the rescaling by using data that has been collected in equilibrium conditions. I explore the use of the dissociation constants (Kd) in the model in order to see if they are unrealistic. The rescaled cL promoter affinities are presented on Table 4.1. For the first time the use of volumetric data now becomes important. Kds are

reported in Molar units. Therefore, without this information I could not judge how realistic these affinities are.

Table 4.1 Dissociation constants for cL binding to its targets after rescaling, in U2017.2. The binding affinities are corrected by the number of Chip-seq sites reported by Nagel et al 2016.

promoter	*U2017.1.2*	*U2017.1.2 corrected*
PRR9	322	0.29
PRR7	597	0.54
PRR5	261	0.23
TOC1	172	0.15
LUX	340	0.30
ELF4	340	0.30
ELF3	369	0.33
GI	500	0.45

The first question to ask is whether these Kds fall within the range of values measured in plants. The affinity for that promoter is 405 nM in U2017.2 (Table 4.1). The reported Kd for CCA1:LHY dimer is 1.44 nM +- 0.2 and around 5 nM for the CCA1 dimmer for a sub-region of the *PRR9* promoter that presents and EE box measured by Surface Plasmon Resonance SPR (O'Neill et al. 2011) I explored the literature building an empirical distribution of dissociation constants (Figure 4.5).

The dynamic range of cL spans $1x10^4$-$1x10^5$ molecules per nucleus. It is important to remark the next expectation for TFs. In order for the clock to be effective, the TF should be able to modulate the transcription rate of its targets. Therefore, in principle their concentrations should vary around the K_ds. Much

higher concentrations would result in continuous activation or repression, or conversely no regulation if the concentration was always much lower than the Kd.

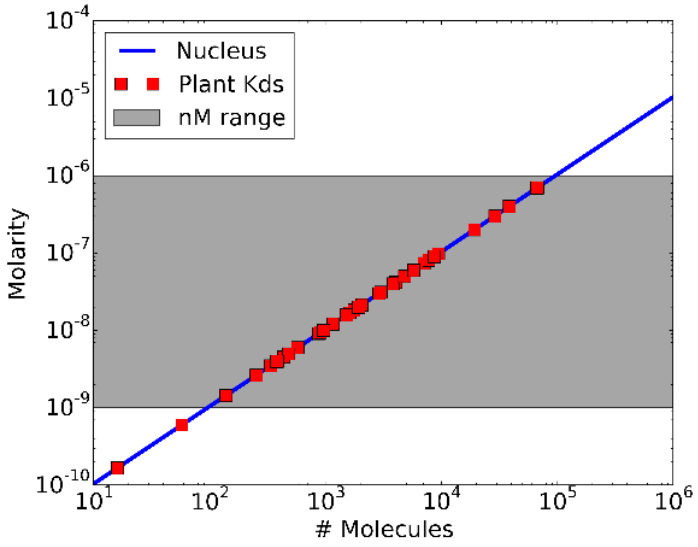


Figure 4.4. K_d distribution measured for Arabidopsis Transcription Factors. Kd as function of protein monomers in the nucleus of Arabidopsis. In red Kds reported in the literature. In grey nanomolar range.

There is a gap of more than two orders of magnitude emerges for the reported CCA1 and LHY affinity for the PRR9 promoter compared to the rescaled U2017.2 (Table 4.1). This result at first sight suggested that the Kds of cL targets in the model are unrealistic. But, it is important to remember that the amount of protein that the plant produces might be covering not only the clock demand but interactions with other molecular components, e.g. other promoters or complexes. Several types of interactions can be imagined. However, I will assume that the main demand for TFs is transcriptional regulation through DNA binding. This could be widespread, as transcriptomic

data has shown that a considerable proportion of genes present circadian rhythmicity (Edwards et al. 2006; Hazen et al. 2009)1.

The SPR data represents the affinity that the Transcription factor experiences for single site. In the genome other sites are present, which requires a higher concentration of TF for regulating them, as each site contributes to the total affinity that the TF experiences for the genome. The Kds in the model originally consider only one site. If the model's simulated protein levels is appropriate for the whole genome, then the affinity for promoter in the model needs to decrease (Kd values drop) so that the dynamics are unaffected. A trivial way to do this is by assuming that each site is independent in the genome. Furthermore, I will normalise the U2017.2 Kds for single site in the genome assuming by dividing the original Kd by the total number of Chip-seq sites. The estimation of Chip-seq peaks in the genome comes from ~1100 (Nagel et al. 2015). A more complex treatment can be taken using statistical mechanics(Simicevic et al. 2013). However nuclear size information and number of Chip-seq sites simplifies greatly this allowing the use of simple chemical kinetics. Table 4.1 shows the K_d correction by the number of sites using Nagel data. The resulting PRR9 sequence K_d is closer to the measured value, now around one fifth of the reported affinity of the CCA1:LHY dimer K_d for that promoter. Nonetheless this suggests that the affinities in the model are biologically reasonable for CCA1/LHY targets.

TF binding predictions, constrained by multiple data types and a biophysical model.

To assess the contribution of other potential binding sequences within the *PRR9* promoter, where no Kd value had been measured, I harnessed the data from Protein Binding Microarrays (PBMs). In this technology, the TF is labelled with a fluorophore and assayed on a DNA chip that contains all possible 8-mers, therefore allowing to study transcription factor sequence preferences in a high-throughput way (Berger & Bulyk 2009). This technology has been applied to several organisms for inferring binding motifs. (Weirauch et al. 2014). PBMs have been reported for two clock TFs, CCA1 and LUX (Helfer et al. 2011; Franco-Zorrilla et al. 2014). In principle PBM signal intensity should be correlated with affinity determined by SPR. If the correlation is linear, a regression can be performed to calibrate the PBM signal. Figure 4.6 shows the regression by using the sequences reported by O'Neill et al.

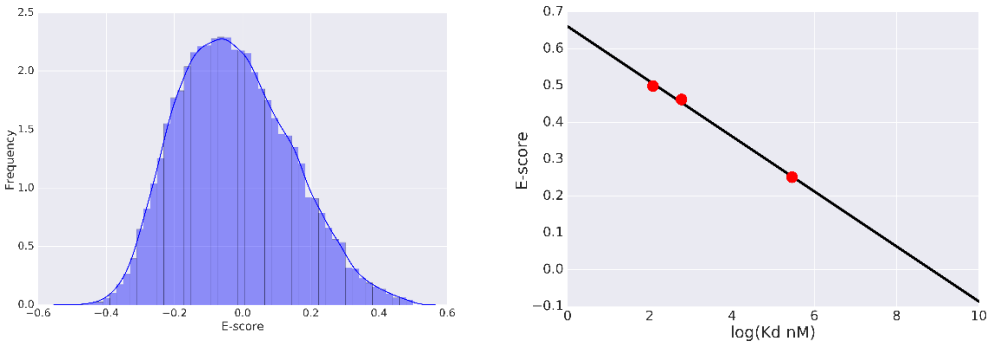


Figure 4.5 Enrichment-score (E-score) distribution of CCA1 PBM. Regression of E-scores against SPR data for CCA1 homodimers Binding to three promoter sequences in (O'Neill et al., 2012).

Figure 4.6 also shows that many low affinity 8-mers will contribute to the total affinity. The lowest affinity site in the SPR data shown in panel B was a mutated sequence with Kd of ~300nM, corresponding to a PBM E-score of ~0.25. I could have used the mutant site as threshold. However, the issue is that the experimental error in PBM signal would not be considered and therefore sites that were unbound could be included as competing with the TF available to the clock promoters, and the other way around (Figure 4.7). The aim here is to estimate the total number of bound sites in the genome, not to find the most reliable, high-affinity sites, in contrast to the most common situation.

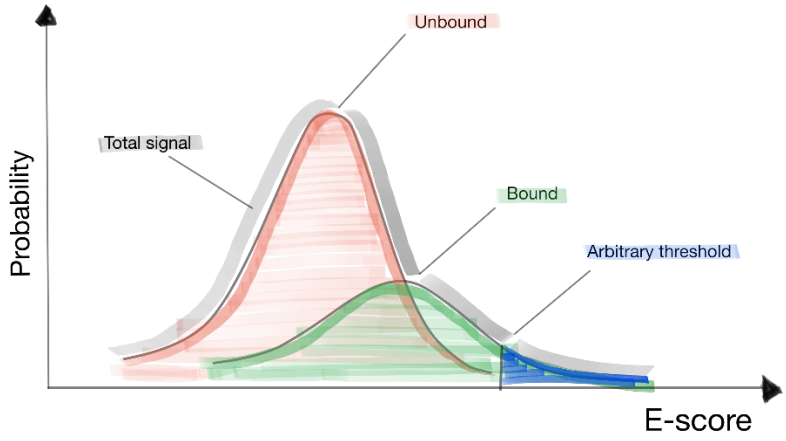


Figure 4.6 Selecting sequences for affinity assignment. In grey, the intensity distribution. An arbitrary threshold could select a subset of real sites that are bound.

Instead of using an arbitrary threshold I decided to infer a biophysical model that would allow me to dissect the intensity distribution in bound and unbound sequences. This biophysical model consisted of an energy matrix (EMA) of

dimension 4x8. Each entry represents the energy contribution of a particular base in the sequence to the total binding energy. The model assumes independence between bases and the total energy is calculated by adding up the individual base contributions. If the energy of a sequence is below a certain energy threshold then it is assumed to be bound by the TF unbound otherwise. The 4x8 parameters for the model need to be inferred from the PBM results. A Bayesian approach can be taken for this. Nonetheless, a likelihood function needs to be defined. A Gaussian experimental error could be used for calculating the likelihood done by the MatrixReduce algorithm (Foat et al. 2006). But it has been pointed out that the way the PBM arrays are processed might compromise the use of this likelihood. Instead an average error model can be used (Kinney et al. 2007).

In this approach, the PBM Enrichment scores (E-scores, *z)* are discretised resulting in the likelihood

$$p(\{z_i\}|\{\sigma_i\},\theta) = \prod_{z,x} p(z|x)^{c_{zx}}$$

Equation 4. 6

Where *p(z|x)* is the error model, indicating the probability of an E-score given that the sequence is bound x=1 or unbound x=0. c_{zx} is the number of

sequences σ_i for which z_i=z and x_i=x. Using an information theory approach Kinney et al derived the following likelihood function

$$p_{ema}(\{z_i\}|\{\sigma\},\theta) = \frac{\Gamma(m)^n \prod_{z,x} \Gamma(c_{zx}+1)}{\prod_x \Gamma(c_x+m)}$$

Equation 4.7

This particular likelihood instance is derived from assuming a flat prior for the error model. Kinney at al also described an MCMC Metropolis Hastings (MCMC MH) sampling scheme for inferring the parameters of the energy matrices.

After implementing these ideas in a python script (see Data Appendix), I applied the approach in the inference of energy matrices for CCA1 and LUX DNA-binding from reported PBMs (Helfer et al. 2011; Franco-Zorrilla et al. 2014). The likelihood chain for the CCA1 matrix reached a likelihood basin after $5x10^3$ steps (Figure 4.8A). Another important thing to verify is that the parameters chains are well mixed. This can be appreciated by plotting the evolution of the parameter values (Figure 3.8B).

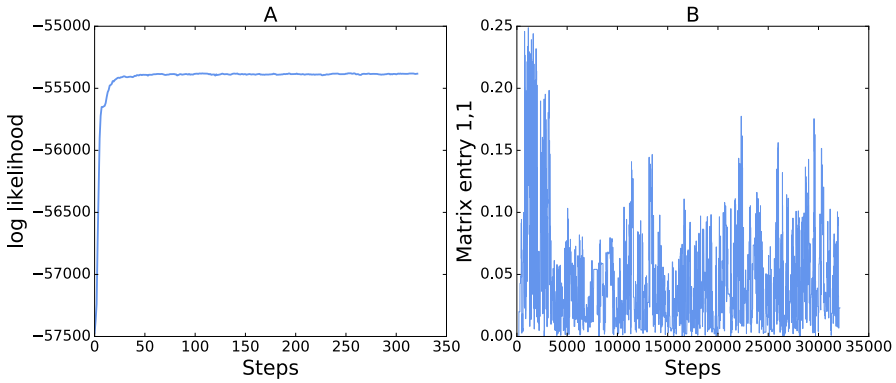


Figure 4.7 CCA1 energy matrix inference. A) MCMC MH log likelihood chain for the ensemble of CCA1 energy matrices, only 1 of every 10 steps plotted. B) Single parameter for CCA1 illustrating the chain behavior for the total ensemble of matrices.

After the inference took place I performed the deconvolution of the E-score values in the PBMs for CCA1 and LUX PBMs (Figure 4.9). The deconvolution shows that the arbitrary threshold E-score could have missed many sequences that could be considered as bound but given the experimental error would have been left out from the energy calculation. This allows me to eliminate the use of the arbitrary threshold and include in the affinity assignment only the sequences considered to be bound by the energy matrix.

In order to judge the use of these matrices I decided to calculate energies for sequences that were observed to be bound by CCA1 and LUX from Chip-seq experiments (Nagel et al. 2015; Ezer et al. 2017). Calculating the binding energy for promoters *in vivo* is still an unsolved question. I am aware that position effects and nucleosome positioning have an important impact on gene expression levels (Sharon et al. 2012). Data about genome accessibility and

nucleosome positioning is available (Sullivan et al. 2014; Zhang et al. 2015). I opted first for a very simple approach, where I just add the contribution of each 8-mer in the sequence to obtain a total affinity for each promoter sequence. I also ignore any possible interference between sites to simply the calculations.

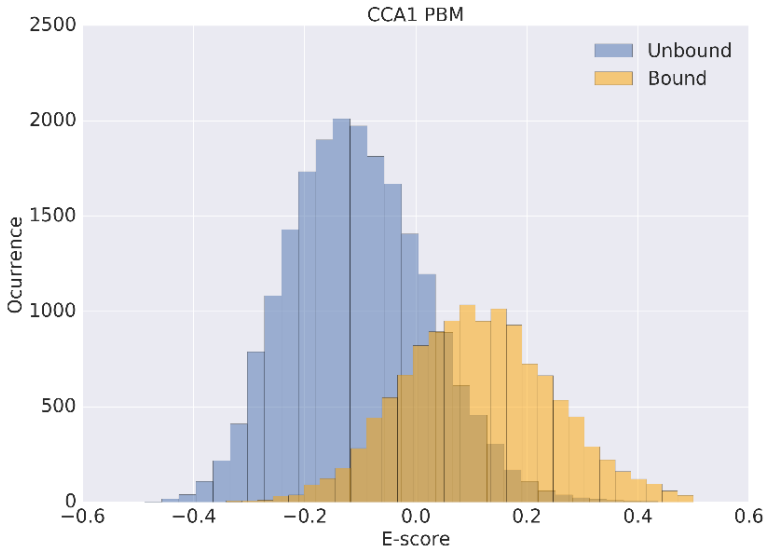


Figure 4.8 De-convoluted E-score for CCA1. Sequences on the PBM were classified by using the inferred energy matrix

Using the boundaries of sites identified from the ChIP-Seq experiments mentioned above I calculated the energies using the TAIR10 Arabidopsis genome annotation. First, I calculated the affinities by using all the PBM data, calling this analysis the Full E-score. The scoring process proceeds by scanning the selected sequences with the energy matrix moving by one nucleotide for each step selecting sequences that might contribute to the total promoter affinity. I assume that interaction between sites is weak. Therefore, I

treat the promoter in a linear way. In many cases the number of binding boxes of the same case tend to be sparse with just a handful of them. I do this for the whole list of bound regions reported by Nagel et al (Figure 4.10A). Notice the change in shape of the distribution after introducing the energy matrix. Many of the ChIP-seq peaks are removed from the "high" affinity region, resulting in a narrower distribution. This suggested that the matrix approach could help to avoid overestimating the affinity of sites identified by ChIP, where the unfiltered PBM data suggested high-affinity binding. The technical implications of this result are outlined in the Discussion.

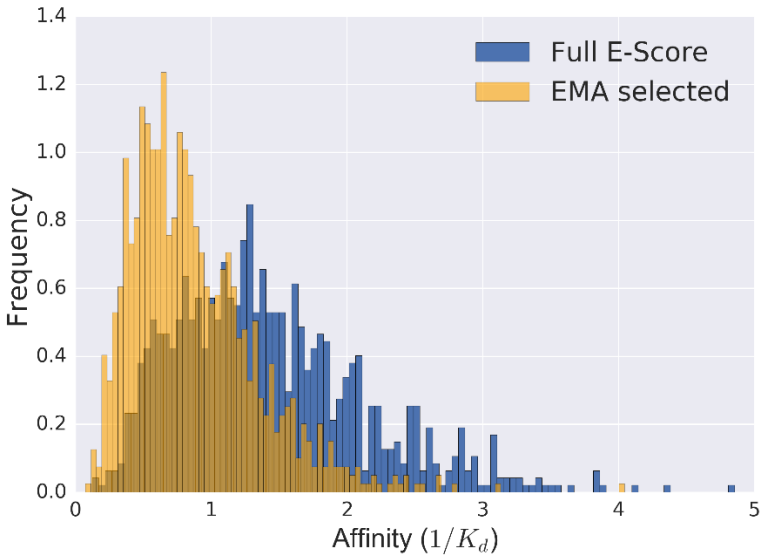


Figure 4.9 Affinity calculations across the genome for CCA1. Blue) calculated affinity using al PBM 8-mers. Orange) Using only sequences selected as bound by EMA. A) Affinity assigned for Chip-seq peaks using TAIR10.

Then I needed a way of assessing whether the inferred affinity distribution could be obtained by chance. I concatenated the Arabidopsis genome and drew from it a set of 1,000 random sequences for each CCA1-bound, ChIP-

seq peak. I calculated the affinity for each of those randomly selected sequences and built a distribution. This random distribution can be approximated by a log-normal model, allowing the calculation of a Z-score for each of the observed ChIP-seq sequences, relative to the distribution of random sequences (Figure 4.11 upper panel). Doing this for all CCA1 ChIP-seq data showed that the observed ChIP-seq sequences were strongly skewed towards higher CCA1-affinity than expected from the random sequences, as expected. However, there was a further increase in significance when CCA1-affinity was predicted using the EMA-filtered data (Figure 4.11 lower panel). Whereas CCA1-affinity for random sequences on the PBM was overestimated by the Full E-score (Figure 4.10B), sequences that were bound *in vivo* in the ChIP-seq data were preferentially scored with higher affinity after EMA selection rather than by the Full set of E-scores. This result suggests that the EMA-selected PBM data offer a reasonable approach to asses binding strength to a sequence of interest. I summarise the assignment by this approach in Table 4.2.

Table 4 2 PBM EMA –selected affinity assignment to clock promoters vs affinities predicted by U2017.2 after rescaling and corrected by number of Chip-seq sites.

promoter	*U2017.2*	*U2017.2 corrected*	*PBM*	*z-score*
PRR9	322	0.29	0.44	2.13
PRR7	597	0.54	0.31	2.23
PRR5	261	0.23	0.47	2.06
TOC1	172	0.15	0.4	2.14
LUX	340	0.30	0.34	2.02
ELF4	340	0.30	0.25	2.7
ELF3	369	0.33	-	-
GI	500	0.45	0.57	1.27

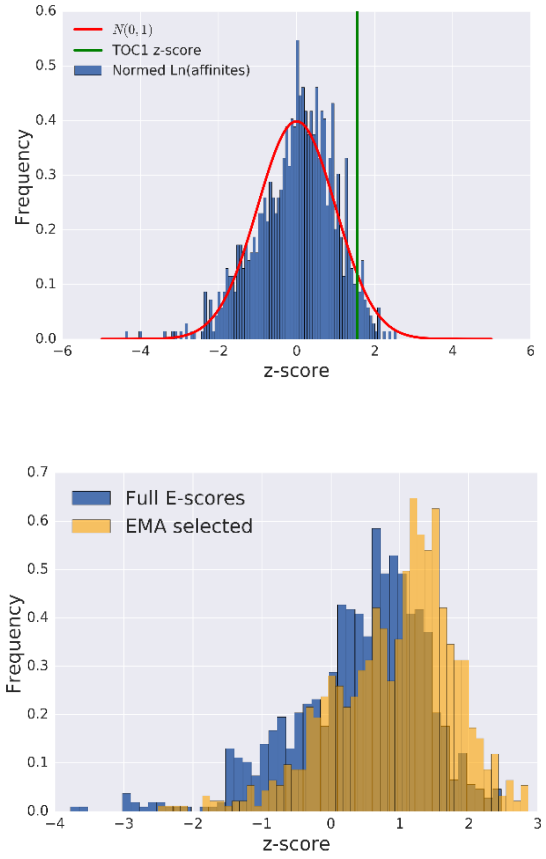


Figure 4.10 The EMA selects sequences that are bound in vivo*. Upper panel) Significance of the affinity of CCA1 for the TOC1 promoter sequence (green), compared to 1,000 random regions from the genome with the same length as the TOC1 ChIP-seq peak. Red, normal distribution fitted to the ln-transformed results for random sequences. Green, affinity for ChIP-seq peak of the TOC1 promoter. Lower panel) Z-scores across all CCA1-bound regions from ChIP-seq data, with affinities assigned with full E-scores (blue) or with EMA selected 8-mers (yellow).*

Then I compare the affinities that U2017.2 presents against the PBM assigned affinity. For the test promoter *PRR9*, I observe only a 0.15 nM difference between the mechanistic model and the matrix assigned affinity. In general, the affinities assigned are not radically different at this scale (Figure 3.12). These results suggest that the predicted affinities of CCA1 on its target

promoters were biologically reasonable, after normalisation to the demand from other sites in the genome. it supports my interpretation that U2017.2 and previously P2011 models considered promoter regions rather than individual binding sites, so affinities in the models cannot be directly compared to the SPR data for individual sites. Finally, note that the affinities derived in U2017.1.2 depend on the protein levels predicted by the simple translation model, which remain to be tested (see Chapter 5). However, it also predicts what the protein levels might be for CCA1 and LHY.

Promoter affinity can explain difference in amplitude between *GI* and *TOC1* RNA waveforms.

Before continuing with the next genes, I wanted to challenge the PBM results against something not intuitive but related to the biology of the clock. In particular the expression dynamics of *GI* and *TOC1* are an interesting pair to compare. Both have similar regulation but the predicted CCA1 affinity differs. Performing work on GI is also relevant, as it is an important regulator of flowering time, therefore variability in its promoter regions might explain some of the observed variability of flowering in ecotypes from different geographical locations (de Montaigu & Coupland 2017).

Interestingly the calculated affinities of cL for *GI* and *TOC1* using EMA resulted in a similar trend as in the U2017.2, with lower affinity for GI compared to TOC1 (Table 4.2). I proceeded to introduce the EMA affinities for *TOC1* and *GI* into

U2017.1.2, observing that the levels for the transcript of these two genes increase. The network is robust to this perturbation, reflected in the persistence of oscillatory behaviour in constant light (Figure 4.13). However, I did not tested correct phenotypes for the rest of the mutants present in TiMet collection. Nonetheless, this suggest that U2017.1.3 can operate with affinities derived from the *in vitro* PBM, SPR measurements and the biophysical EMA. Overall, the integration of several data sets for CCA1 promoter binding suggest that the predicted levels for CCA1 and LHY monomers from the simple model are reasonable, and provide strong foundations for interpreting experimental work when quantifying these proteins in Chapter 5.

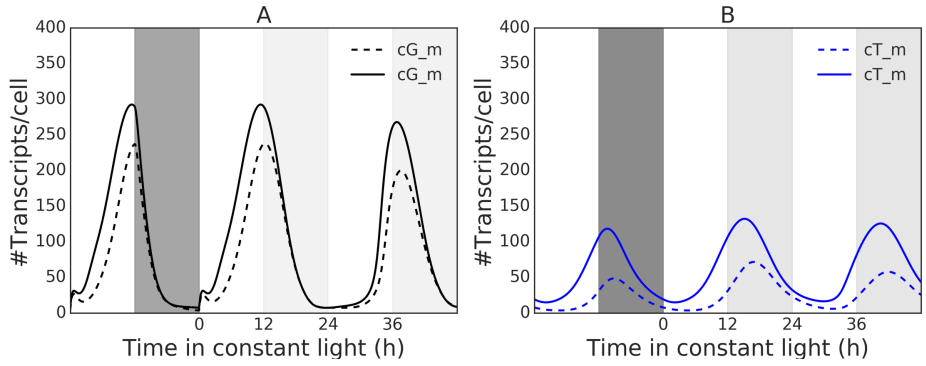


Figure 4.11 Effect of EMA affinities in U2017. A) G transcript I dynamics in the U2017.2 (dotted lines) and when the EMA-selected PBM affinities are introduced into the model (solid lines)

Rescaling PRR translation suggests unrealistic rates.

The regulatory effect of the PRR proteins on the promoter of *LHY/CCA1* (cL_m) is represented as a simple sum of the protein concentrations

$$n\frac{g^2}{g^2+(cP9+cP7+cP5+cT)^2}$$

Equation 4. 8

n is the light independent transcription rate of cLm, g is the affinity the PRRs experience for the promoter of cLm. This simple expression nonetheless enforces a constraint on the scaling factor. The variables representing the PRR proteins must have the same units before the sum in the denominator can be performed. To illustrate this point, consider rescaling any of these variables, for example *cP7*, with a scaling factor *γ.* Let *cP7' = γ. cP7*, then

$$n\frac{\gamma^2 g^2}{\gamma^2 g^2+(\gamma cP9+\gamma cP7'+\gamma cP5+\gamma cT)^2}$$

Equation 4.9

Observe that the scale factor applies also to the affinity *g*, representing a change in affinity. This change in affinity is the result of compensating for the change of *cP7* scale, similar to the cL rescaling above. Notice that performing the rescaling also rescales the other variables by the same factor. This also shows that determining the absolute levels of one of the PRRs is enough to predict the levels for the other PRRs. Similar to cL, I rescaled the PRRs using PRR7 as the reference protein (Figure 4.15). During simulated LD, cP7 and cT both closely matched the levels in the SM training data (Figure 4.14). The peak of cP7 is 3 hours earlier than the synthetic data profile (SDP) in LD.

Observe also the difference in decay rate between the LD and LL regimes for cP7 in U2017.2 and less change in the SDP.

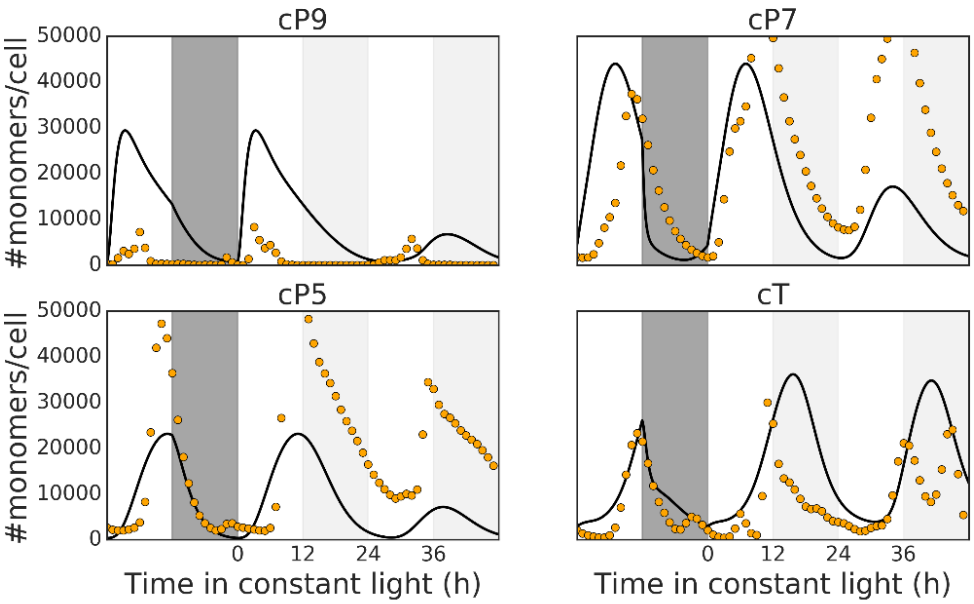


Figure 4.12 Rescaling PRRs in U2017. In solid black U2017.2 matching synthetic data profile (SPD) from the SM. Synthetic data in yellow circles. The model was entrained to 10 LD cycles and then released into LL.

I decided to verify the translation rates in U2017.2 and compare them to the translation rates assumed in the simple model of translation (SM). I did this by taking the product of translation rate used from the SM and the mRNA level in U2017.2. Then I obtained the translation rates in U2017.3 which has rescaled protein for matching the predictions of the SM (Figure 4.15). I can notice immediately that the translation rates of cP7 and cP5 are higher than would be expected from the translation constant used for generating the SDP. This is a consequence of the model having a short half-life of around one hour

in constant light for both cP7 and cP5. Correspondingly high translation rates are required to produce sufficient protein to regulate cLm. One possibility for solving this issue would be reducing the degradation rates. This has an important consequence for the peak of expression of cLm in the first day in constant light. Already the decay of the PRRs is two slow in this condition, delaying the peak several hours after expected dawn, significantly later than observed in RNA data (Figure 4.16). PRR7 protein has a very strong role in the regulation of cLm and it is enough for strong repression of cLm in U2017.2 (red trace in Figure 4.15). Therefore, decreasing the degradation rates would result in a slower disappearance of the PRRs producing an even later peak of cLm during the first day of constant light. This leads to a paradoxical situation, where unrealistically high degradation and therefore translation rates trade off against the delayed rise of cLm in LL.

A new mechanism should be incorporated to resolve this issue. Furthermore, It is possible that the simple model underestimated the PRRs levels. As many as 6,427 sites have been reported from Chip-seq experiments (Liu et al. 2016). Assuming that all PRRs are equivalent in the genome and correcting the affinity by this number of sites results in a predicted PRRs Kd for cLm of 0.0129 nM in U2017.2. This is more than two orders of magnitude higher than the reported affinity of CCA1, which is surprising and suggests that the number of bound sites was over-estimated or the active protein level was under-estimated. Unfortunately, no affinity data for the PRR proteins has been

reported to verify these numbers. Measurements of protein levels might help for gaining insight in this issue.

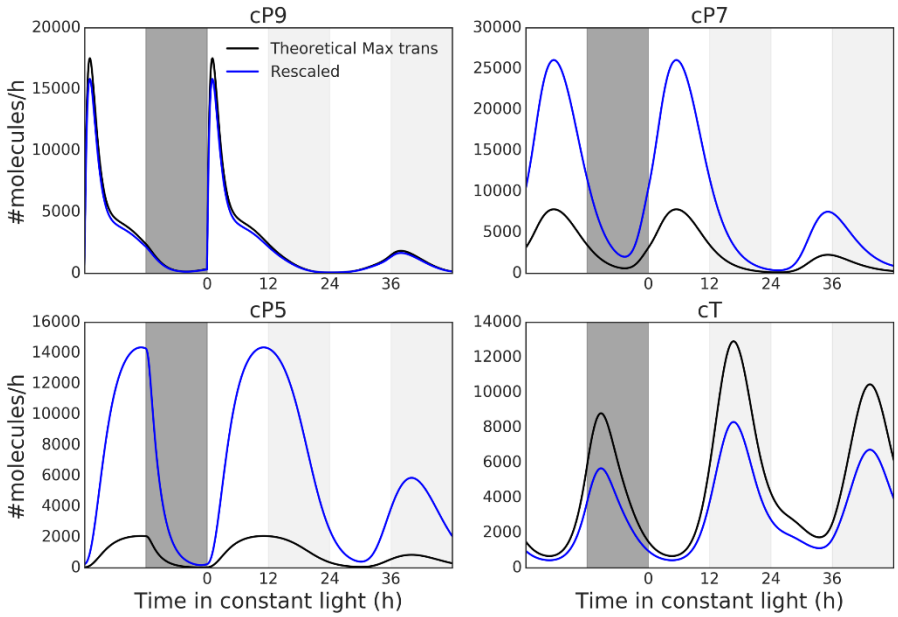


Figure 4.13 Translation rates of PRRs in U2017 are unrealistic. Translation rates using the predicted RNA from U2017 Rates obtained by assuming the same translation constant as for the simple mode. Translation rates after rescaling protein variables matching PRR7 levels. PRR7 and PRR5 show the strongest disagreement.

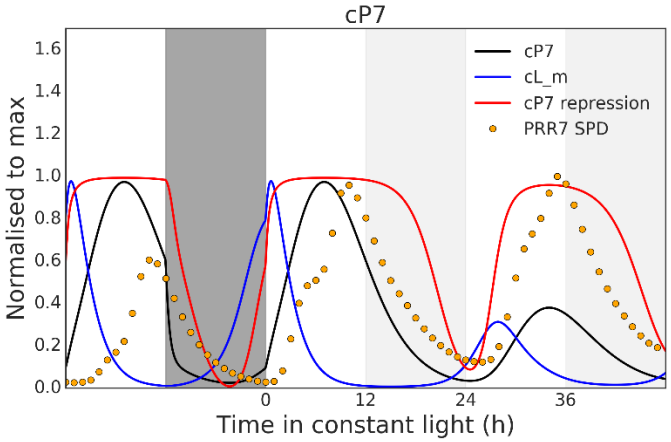


Figure 4.14 Late rise of cLm in U2017.2. The dynamics of cP7 in U2017.2 (Black), cLm in (Blue). cP7 Synthetic protein data (SPD) (Dots). Repression of cLm expression assuming only cP7 as regulator in U2017.2 (Red). All traces normalized to maximum.

Rescaling the Evening Complex

LUX, ELF3 and ELF4 proteins form the Evening Complex in our clock models. ELF3 protein also interacts with GI, ZTL and COP1 proteins, affecting protein degradation rates. Each of these components in turn has multiple interactors. Rescaling these variables is a subtler process than for cL and the PRRs. First, I inferred decay constants for predicting protein levels by matching the simple data driven model to western blot data (Figure 4.19). The fitting equation 4.1 and the scaling factors were automatically obtained by using equation 4.2. The mas scale was fixed as described for cL and PRRS.

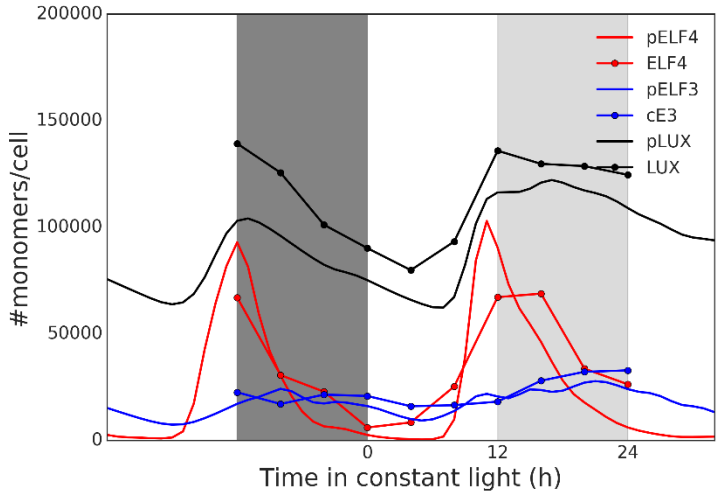


Figure 4.15 Predicting EC components' abundance using TiMet data. Cubic interpolated RNA data were input to the simple model described above. Initial conditions and degradation rates were fitted to figures in (Nusinow, Helfer, Hamilton, King, Imaizumi, Schultz, Farré & Kay 2011a). Data were rescaled using an analytically determined scale factor B.

Any of the components of the Evening Complex could be measured to achieve this rescaling, if the stoichiometry of complex formation would constrain the levels of all the interacting partners. In particular, I found by inspecting the equations that the cytosolic concentration of ELF3 could be used to rescale all the EC components and sub-pools. But from an experimental perspective, this variable is problematic as ELF3 is mainly present in the nucleus (Herrero et al. 2012). Therefore, measuring *cE3c* would require fractionation of cell compartments, which is a very time consuming and delicate process that might significantly impact the experimental error.

The LUX protein variable *cLUX* is a good alternative. It has a simple representation in U2017.2. Relevant PBM, Chip-seq and affinity data exists

(Helfer et al. 2011; Silva et al. 2016; Ezer et al. 2017). The problem with LUX is that it can fix the mass scale for the EC components but not for other interactors in the model. Circumventing this issue is possible by rescaling the central cE3c variable but performing the costing against cLUX. Figure 4.20 shows the results of this approach using synthetic protein data and fitting against cLUX protein predictions.

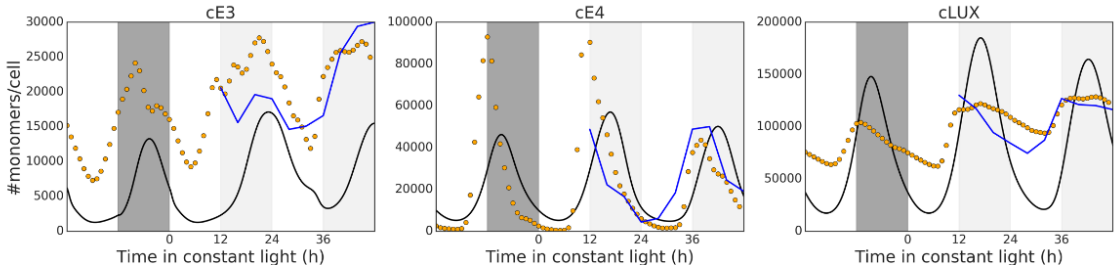


Figure 4.16 Rescaling the components of the EC. U2017.2 Evening Complex components where rescaled against cLUX SPD. Orange) represent synthetic protein data form the simple model. Blue) is a rough profile of the Evening Complex components obtained by processing Figure 2 from (Nusinow, Helfer, Hamilton, King, Imaizumi, Schultz, Farré & Kay 2011b) using imgagJ. Black) U2017.2 rescaled to match the LUX protein scale.

The Evening complex presents unrealistic affinities for target promoters.

The rescaling of the EC in U2017 using LUX synthetic data uncovered that the binding affinities are unrealistic. The nominal affinities in the model fall in the sub-nanomolar range which can be considered to be biologically reasonable (Table 2). Furthermore, LUX binds to more than 800 genomic regions (Ezer

et al. 2017). Accounting for these sites results in affinities in the pM range. These values are very unlikely to be observed in plants (Figure 4.5). The predicted pM affinities are the result of a very late peak of EC activity in the middle of the night. This activity is needed at this point because it is required for repressing the PRRs at the middle of the night. However, this contrasts with the E.C. profile obtained by Co-IP (Nusinow, Helfer, Hamilton, King, Imaizumi, Schultz, Farré & Kay 2011a). At the start of the night, the simulated EC has reached only a small fraction of its peak abundance but this is its required time of action on the evening genes (Figure 4.12). The simulated E.C. must therefore have a very high affinity, to be active at a low complex concentration.

Table 4.3 Rescaled affinities of the EC in U2017 for four target genes

promoter	*U2017.2*	*U2017.2 corrected*	*PBM*	*z-score*
PRR9	9.25	0.012	0.27	1.94
TOC1	10.2	0.013	0.68	1
ELF4	8.85	0.011	0.25	2.7
GI	3.58	0.004	-	-
LUX	-	-	0.26	1.43
PRR5	-	-	1.97	1.1
CCA1	-	-	0.39	0.66

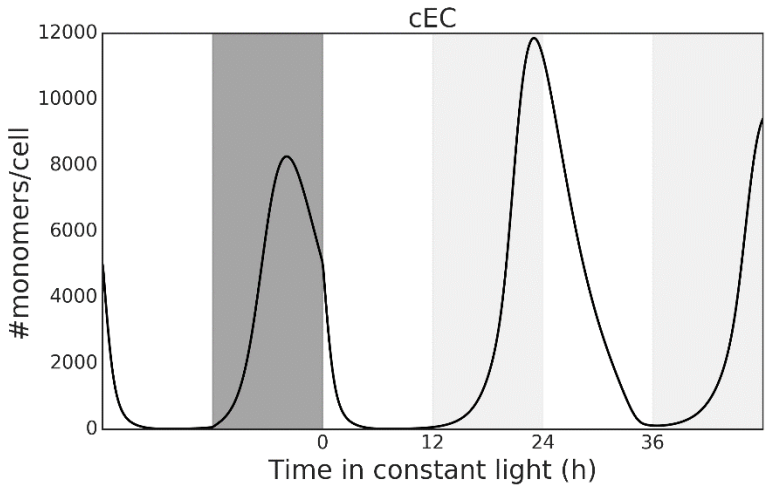


Figure 4.17 Evening Complex dynamics in U2017.2. The evening complex peaks late during the night. The U2017.2 model was entrained to 10xLD cycles and released into constant light conditions at time 0.

I also performed affinity analysis for the promoters that are regulated by LUX, similar to CCA1. Only one dissociation constant has been determined for LUX (Silva et al. 2016). Therefore, a regression cannot be performed. Therefore, I assume the same slope inferred from the CCA1 PBM-SPR regression. Fortunately, while performing this analysis Chip-seq data for the three canonical members of the Evening Complex (LUX, ELF3 and ELF4) was published. Two LUX Chip-seq experiments were reported at 17ºC and 22ºC. More targets were detected at 17ºC. In this set I found *TOC1* but not *GI* and in 22ºC the opposite happened. The z-scores are weak compared to the genome background, all of them predict sub-nM Kds apart from PRR5 (Table 4.3). In the model before correcting for genomic demand, the predicted affinities were in the low (1-10) nM. I want to point something interesting. The rank order of PBM results are consistent with the U2017.2 affinities. The EC shows affinity

for GI > LUX > PRR9 > TOC1 in U2017.2. Which is the same order in the PBM results apart from *GI* which was not detected at 17ºC. Nonetheless, Ezer and co-workers detected both *GI* and *LUX* promoters at 22ºC and not any of the other components of the clock.

Discussion

I proposed to make predictions about the levels of clock proteins in order to understand how the use of absolute units could challenge the model. Performing this theoretical exercise is remarkably beneficial. Several relevant results emerged that will be important when assessing the model when I perform the absolute quantifications for the selected proteins in this chapter. I showed that genomic data of binding is required to properly compare the inferred model affinities from time series and in vitro quantifications of binding. I showed that SPR data can be used for calibrating Protein Binding Microarrays. The calibrated microarrays allow assignment of affinity to promoters. This result closed the gap between SPR experimental data and the U2017.3 affinities after being corrected by number of Chip-seq peaks in the genome. This result suggests that U2017.3 reports promoter affinities rather than affinities for single binding elements on the promoter of clock genes. I introduced the use of a machine learning approach for deconvoluting the PBM distribution into sequences that are more likely to be bound by CCA1 and LUX. Avoiding in this way artificially underestimation of affinities by choosing an arbitrary threshold of binding in the PBMs.

A concern related to the PBMs is whether the amount of protein used for CCA1 was saturating. However, verifying in detailed the experimental procedures reported showed the following (Godoy et al. 2011; Franco-Zorrilla et al. 2014). The protein purified was MBP-CCA1 used a was 1μg in 100μl of binding buffer. The size of MBP-CCA1 is around 109.975 kDa. Resulting in a concentration of 9.13 nM in the PBM conditions. The reported Kd for CCA1 dimmer for CCA1 best binding site is ~2.5 nM (O'Neill et al. 2011). Therefore, under the PBM binding conditions and assuming a hill function with h=2 there is an occupancy probability of more than 90%. Other experimentally tested sites of CCA1 present reduced affinities which are below saturation under the PBM binding conditions (Figure 4.18).

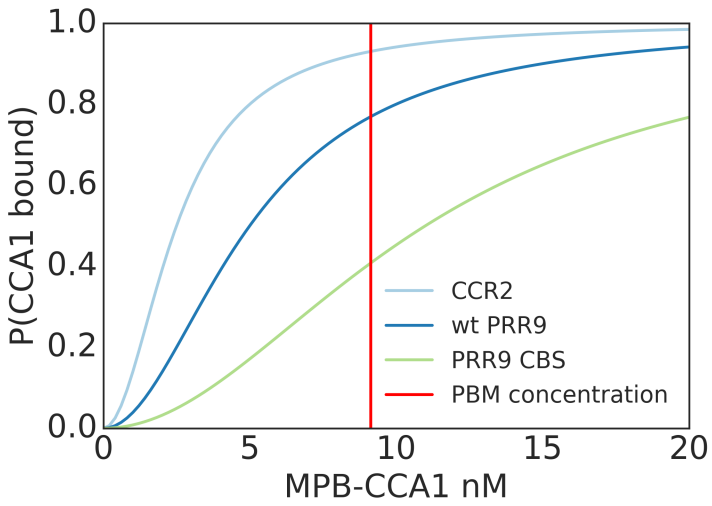


Figure 4.18 Binding probability of CCA1 for different sequences with experimentally determined Kds from (O'Neill et al. 2011)

This analysis could be extended to LUX which presents an affinity around 36 nM (Silva et al. 2016). Suggesting that the LUX PBM is not saturated either. However, I need to remark that Silva et al only reported the affinity of the DNA binding domain instead of the full protein. Therefore, the LUX results should be taken carefully. Nonetheless the predicted affinities are in a similar magnitude as for CCA1.

Revising the representation of the EC

As I mentioned in Chapter 3 it is possible that the evening complex needs to be dissected into two activities across the evening and late night. Performing this operation will allow EC affinities that are within biologically supported values. It has been suggested that *BROTHER OF LUX ARRHYTHMO (BOA)* could cover this role. F2014 model introduced this element in the equations for representing the Evening complex (Fogelmark & Troein 2014). However, the complexity results seem to have resulted in over fitting of this model. The authors reported eight parameter sets. Overexpression of *BOA* results in a log period phenotype (Dai et al. 2011). Surprisingly each of the parameters produce very different period predictions when overexpression is simulated (Figure 4.19). This result means that more data for BOA is needed that will be useful for modelling purposes. In Chapter 3 I showed that dynamics of BOA-LUC presented peak of expression towards the end of the night. I tested this lines in combination with a CCA1p:LUC in plate reader experiment as described in the methods Chapter (Figure 3.20).

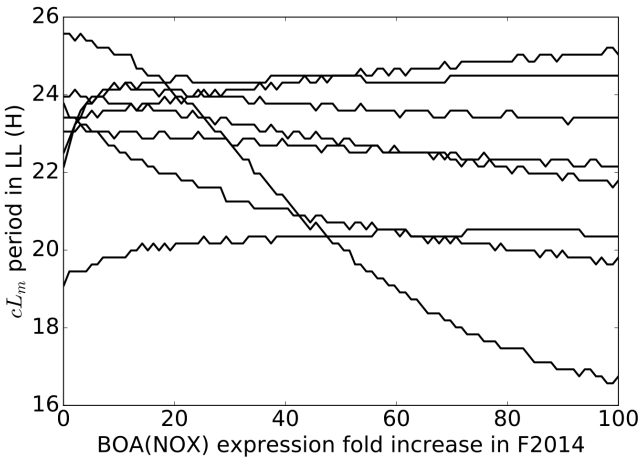


Figure 4.19. Impact of overexpression on the period of cLm in constant light in F2014.

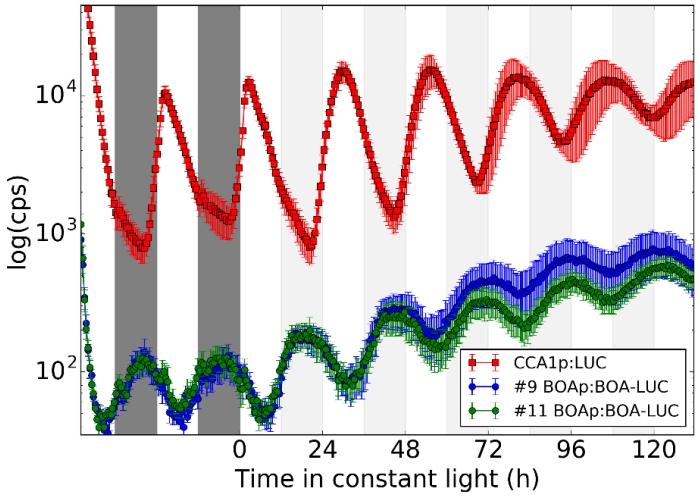


Figure 4.20 Dynamics of BOAp:BOA-LUC T2 line in diurnal and constant light conditions. Mean of 4 biological replicates error bars S.E.M. 2x12L:12D cycles and then transferred into constant light conditions. In red, Col-0 CCA1p:LUC as control. Two representative T3 BOA lines #9 and #11 plotted, blue and green.

The results confirmed the observations on T1 lines, performed with 2-week old plants. The peak of fusion protein signal occurs at ZT21. The LUC signals at peak levels are also an order of magnitude lower than for CCA1p:LUC. Also, oscillations persist in constant light conditions in an anti-phasic relationship to CCA1p:LUC. Reported western-blot data show very weak levels of BOA protein using a specific antibody, which would suggest that BOA is expressed at very low levels *in-planta* (Dai et al. 2011). Which might explain why BOA is not enough to act as a redundant activity for LUX. Further evidence for LUX acting as a component of the Evening Complex includes lower occupancy of ELF3 protein on the promoter region of *PRR9* when knocking down the levels of *LUX* and *BOA* transcripts by using amiRNA compared to the single LUX knock down (Chow et al. 2012). Furthermore, LUX and BOA can be detected by affinity purification mass spectrometry with a tagged version of ELF3 at ZT12 (Huang et al. 2016). I want to propose Taken together, these results suggest BOA levels might not be enough to maintain. However, no null mutants of BOA exist and work into this direction should be done in future work using CRISPR technology. The affinity results suggest that it will be beneficial for the model to include this component in future versions. Also, starting with a simpler model like U2017.3 would help to model over-fitting and properly asses the role of BOA in the clock.

References

Berger, M.F. & Bulyk, M.L., 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols*, 4(3), pp.393–411.

Brown, K. & Sethna, J., 2003. Statistical mechanical approaches to models with many poorly known parameters. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 68(2 Pt 1), p.21904.

Chow, B.Y. et al., 2012. ELF3 recruitment to the PRR9 promoter requires other Evening Complex members in the Arabidopsis circadian clock. *Plant signaling \& behavior*, 7(2), pp.170–173.

Dai, S. et al., 2011. BROTHER OF LUX ARRHYTHMO is a component of the Arabidopsis circadian clock. *The Plant cell*, 23(3), pp.961–972.

de Montaigu, A. & Coupland, G., 2017. The timing of GIGANTEA expression during day/night cycles varies with the geographical origin of Arabidopsis accessions. *Plant signaling \& behavior*, 12(7), p.0.

Edwards, K. et al., 2006. FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. *The Plant cell*, 18(3), pp.639–650.

Ezer, D. et al., 2017. The evening complex coordinates environmental and endogenous signals in Arabidopsis. *Nature Plants*, 3(7), p.17087.

Flis, A. et al., 2015. Defining the robust behaviour of the plant clock gene circuit with absolute RNA timeseries and open infrastructure. *Open biology*, 5(10), p.150042.

Flis, A. et al., 2016. Photoperiod-dependent changes in the phase of core clock transcripts and global transcriptional outputs at dawn and dusk in Arabidopsis. *Plant, cell \& environment*, 39(9), pp.1955–1981.

Foat, B.C., Morozov, A.V. & Bussemaker, H.J., 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics (Oxford, England)*, 22(14), pp.e141–9.

Fogelmark, K. & Troein, C., 2014. Rethinking transcriptional activation in the Arabidopsis circadian clock. *PLoS computational biology*, 10(7), p.e1003705.

Forger, D. & Peskin, C., 2003. A detailed predictive model of the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25), pp.14806–14811.

Franco-Zorrilla, J.M. et al., 2014. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences of the United States of America*, 111(6), pp.2367–2372.

Godoy, M. et al., 2011. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *The Plant Journal*, 66(4), pp.700–711.

Hazen, S.P. et al., 2009. Exploring the transcriptional landscape of plant circadian rhythms using genome tiling arrays. *Genome biology*, 10(2), p.R17.

Helfer, A. et al., 2011. LUX ARRHYTHMO encodes a nighttime repressor of circadian gene expression in the Arabidopsis core clock. *Current biology : CB*, 21(2), pp.126–133.

Herrero, E. et al., 2012. EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the Arabidopsis circadian clock. *The Plant cell*, 24(2), pp.428–443.

Huang, H. et al., 2016. Identification of Evening Complex Associated Proteins in Arabidopsis by Affinity Purification and Mass Spectrometry. *Molecular & cellular proteomics : MCP*, 15(1), pp.201–217.

Ito, S. et al., 2007. Rhythmic and Light-Inducible Appearance of Clock-Associated Pseudo-Response Regulator Protein PRR9 Through Programmed Degradation in the Dark in Arabidopsis thaliana. *Plant \& cell physiology*, 48(11), pp.1644–1651.

Kiba, T. et al., 2007. Targeted degradation of PSEUDO-RESPONSE REGULATOR5 by an SCFZTL complex regulates clock function and photomorphogenesis in Arabidopsis thaliana. *The Plant cell*, 19(8), pp.2516–2530.

Kim, J. & Forger, D., 2012. A mechanism for robust circadian timekeeping via stoichiometric balance. *Molecular systems biology*, 8, p.630.

Kinney, J.B., Tkacik, G. & Callan, C.G., 2007. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2), pp.501–506.

Lee, C. et al., 2001. Posttranslational mechanisms regulate the mammalian circadian clock. *Cell*, 107(7), pp.855–867.

Liu, T.L. et al., 2016. A G-Box-Like Motif Is Necessary for Transcriptional Regulation by Circadian Pseudo-Response Regulators in Arabidopsis. *Plant physiology*, 170(1), pp.528–539.

Lu, S. et al., 2009. CIRCADIAN CLOCK ASSOCIATED1 and LATE ELONGATED HYPOCOTYL function synergistically in the circadian clock of Arabidopsis. *Plant physiology*, 150(2), pp.834–843.

Más, P. et al., 2003. Targeted degradation of TOC1 by ZTL modulates circadian function in Arabidopsis thaliana. *Nature*, 426(6966), pp.567–570.

Merrow, M.W., Garceau, N.Y. & Dunlap, J.C., 1997. Dissection of a circadian oscillation into discrete domains. *Proceedings of the National Academy of Sciences*, 94(8), pp.3877–3882.

Nagel, D.H. et al., 2015. Genome-wide identification of CCA1 targets uncovers an expanded clock network in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34), pp.E4802–10.

Narumi, R. et al., 2016. Mass spectrometry-based absolute quantification reveals rhythmic variation of mouse circadian clock proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24), pp.E3461–7.

Nusinow, D., Helfer, A., Hamilton, E., King, J., Imaizumi, T., Schultz, T., Farré, E. & Kay, S., 2011a. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

Nusinow, D., Helfer, A., Hamilton, E., King, J., Imaizumi, T., Schultz, T., Farré, E. & Kay, S., 2011b. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

O'Neill, J. et al., 2011. Circadian clock parameter measurement: characterization of clock transcription factors using surface plasmon resonance. *Journal of biological rhythms*, 26(2), pp.91–98.

Piques, M. et al., 2009. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. *Molecular systems biology*, 5, p.314.

Sharon, E. et al., 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6), pp.521–530.

Sharrock, R.A. & Clack, T., 2002. Patterns of expression and normalized levels of the five Arabidopsis phytochromes. *Plant physiology*, 130(1), pp.442–456.

Silva, C.S. et al., 2016. The Myb domain of LUX ARRHYTHMO in complex with DNA: expression, purification and crystallization. *Acta crystallographica. Section F, Structural biology communications*, 72(Pt 5), pp.356–361.

Simicevic, J. et al., 2013. Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nature methods*, 10(6), pp.570–576.

Song, H.-R. & Carré, I.A., 2005. DET1 regulates the proteasomal degradation of LHY, a component of the Arabidopsis circadian clock. *Plant Molecular Biology*, 57(5), pp.761–771.

Sullivan, A.M. et al., 2014. Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell reports*, 8(6), pp.2015–2030.

Weirauch, M.T. et al., 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6), pp.1431–1443.

Zhang, T., Zhang, W. & Jiang, J., 2015. Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants. *Plant physiology*, 168(4), pp.1406–1416.

Chapter 5
NanoLUC tools for absolute quantification of clock proteins

In Chapter 3, I presented a theoretical framework for the introduction of absolute units for the protein levels in the circadian clock model. Also, I explored the consequences of the predicted levels for the model's operation. I based the predictions on protein data from various sources, noting that there is a lack of protein data with quality equivalent to qRT-PCR for RNA. Data on Arabidopsis mainly comprises determination of RNA levels by methods such as Northern blotting, RT-PCR or qRT-PCR. Only a small proportion of time-series present in the literature for the *Arabidopsis* clock are protein data. In the collection of time-series used for inferring F2014 parameters only 11% of the time series correspond to protein, all of them lacking absolute units. (Fogelmark & Troein 2014).

Generating time-series of clock transcription factor (TF) protein levels currently is based strongly on western blots. While the first produced protein time series were against native proteins, now the preferred approach is by generating translational fusions of the TF of interest with a general epitope (HA, Flag or G/YFP tags). This methodology has been applied to several major components of the oscillator e.g. PRRs, TOC1, LHY, CCA1, LUX and ELF3 (Nagel et al. 2015; Herrero et al. 2012; Fujiwara et al. 2008; Mas et al. 2003).

These lines have facilitated cellular and tissue localisation studies. However, no report of their use for quantitative studies has been reported up to date. Using the current tagged lines present several drawbacks, including, manual and destructive sampling, and dealing with signal saturation when performing Western blots.

Therefore, I explored the development of a new methodology that could allow absolute quantification of clock proteins and automated tracking of protein rhythms compatible with current firefly luciferase approaches. The methodology is based on the luciferase NanoLUC (NL) (Hall et al. 2012). In principle, the use of this enzyme as a reporter for protein levels should be compatible with equipment that is already present in many circadian labs. I moved away from FLUC into NL, as it presents many major advantages for reporting protein levels. FLUC has a very unstable activity, which is ideal for reporting newly translated protein and transcriptional dynamics, however is problematic for reporting the actual of protein *in vivo*. Nonetheless it has been used for determining CCA1 and TOC1 protein levels in *Ostreococcus tauri* by manual sampling and protein extraction which can help to avoid product inhibition in the FLUC reaction (Corellou et al. 2009).

NanoLUC is a small protein of 19.2 kDa from the deep-sea shrimp *Oplophorus gracilirostris*. It was engineered for high stability ($t_{1/2}$=11.5 days at 37ºC) and codon optimised for expression in mammalian cells (Hall et al. 2012).

Conveniently, the expected Codon Adaptation Index for expression in *Arabidopsis* is remarkably good eCAI = 0.737 ($p<0.05$) calculated by E-CAI (Puigbò et al. 2008), on a scale where 1 is a sequence with the most common codons in the organism. The enzyme only requires O_2 and substrate in the reaction. The NanoLUC substrate is Furimazine, a coelentarzine analogue, that presents a higher stability than luciferin. It is therefore possible that I will be able to detect signal for extended time periods. Also, the enzyme shows stable light emission around 21ºC and stable activity in the physiological pH for Arabidopsis (7.1 in nucleus, 7.2 in cytosol (Shen et al. 2013)). This contrasts with LUC activity which is sensitive to these temperature and pH ranges. Also, LUC suffers product inhibition, resulting in a 'dark' (inactive) pool of enzyme after it is exposed to luciferin (Millar et al. 1992). LUC requires ATP and Mg^{++}, which might fluctuate in a circadian manner (Feeney et al. 2016), whereas NanoLUC does not. If the activity of LUC is determined in plant extracts this might not represent a problem, as the first moment the enzyme contacts the substrate is during the enzymatic assay where all the cofactors can be added in excess. However, these LUC properties compromise the determination of protein levels *in vivo*, which is one aim of this work. Finally, NanoLUC is reported to give orders of magnitude brighter signal than LUC (Hall et al. 2012), therefore in principle lower number of molecules should be detectable. No results had been reported on the application of NL in plant systems or for chronobiology in other organisms. Before using this reporter to

monitor clock proteins, a set of genetic tools and new protocols need to be developed.

Results

Activity and purification of NanoLUC variants

In order to verify the activity of the NL, I cloned the sequence provided by Promega in pNL1.1 into pET28a(+) and performed an induction experiment in *E. coli* BL21 Rosetta2 with 1mM IPTG. After 6 hours of induction at 30 ºC, cells were harvested by centrifugation, re-suspended in Phosphate Buffered Saline (PBS) and 1μl of furimazine (NL substrate, stock concentration) added. Cells carrying an empty pET28a(+) vector did not present appreciable signal (Figure 5.1).

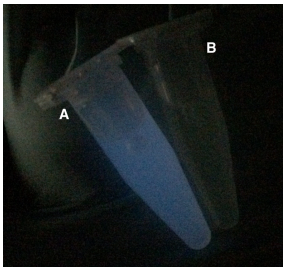


Figure 5.1 NanoLUC expression in E.coli BL21 Rosetta 2. A) Expression was induced for 6 hours at 30ºC with 1mM IPTG. B) Empty pET28a(+) as negative control.

Two derivatives were cloned for future work. A synthetic 3xFlag 10xHis tag was added on the C-terminus of NanoLUC in pET28 resulting in NL3F10H. Addition of the 10xHis allows purification by Ion Metal Affinity Purification (IMAC) using a Ni-NTA matrix as a stationary phase. Also, an MBP-NL-3F10H

(Maltose Binding Protein, MBP) was created to test the effect of fusing proteins to the N-terminus of NLUC. After the MBP, a 6xHis tag was also added, so this NLUC presents 16xHis (sequence in Appendix). The resulting constructs were then introduced into *E. coli* BL21 Rosetta 2 pLysS for expression and purification, as described in the methods chapter. The three versions NLUC, NL3F10H and MBP-NL3F10H presented high expression levels at 30ºC and remained soluble after a centrifugation for 20 min a 20,000xg (Figure 5.2, lysate lane). Then, I performed an IMAC purification using Ni-NTA agarose beads, with a standard protocol. As expected the NLUC original variant does not bind to Ni-NTA beads and can be removed after two washes with Washing buffer (WB) (Figure 5.2A). The NL3F10H version did bind to the Ni-NTA beads and eluted at high Imidazole concentrations present in the Elution Buffer (EB) (Figure 5.2B). The tag allows purification of a homogenous preparation that can be used as a standard. As expected for MBP-NL3F10H a very significant band shift is observed and a homogenous purification can be observed in the NuPAGE gels (Figure 5.2C). This shows that purification of NLUC can be performed in a straight forward way for use in calibration curves.

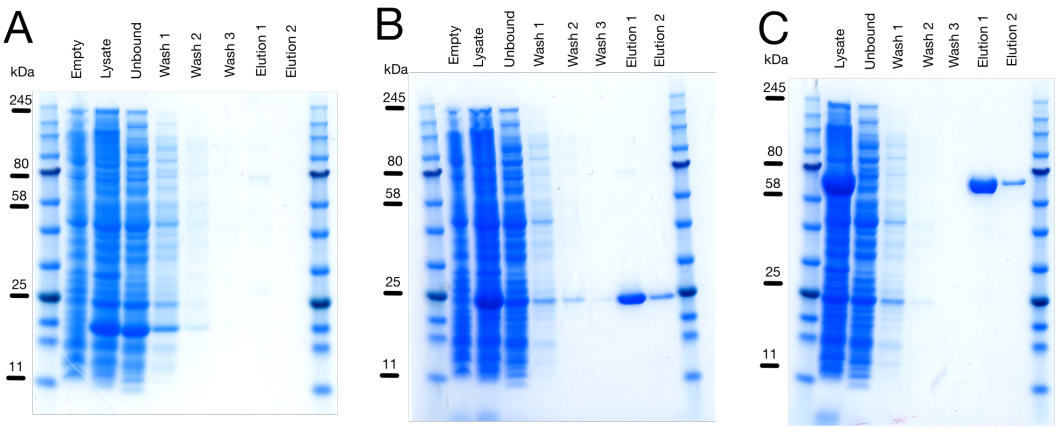


Figure 5.2 Analysis of NLUC variants purification by NuPAGE. A) NLUC B) NL3F10H C) MPB-NL3F10H 10 µl of each fraction were loaded into a NuPAGE 4-12% Bis-Tris and run at 200V for 40 min. Each lane is a step in the purification process.

Then I compared the activity of NL3F10 to MBP-NL3F10H in order to test for possible effects of adding this N-terminal fusion. This is important as I will be using NL3F10H as C-terminal tag for reporting the levels of transcription factors in stable transgenic plants.

I purified NL3F10H and MPB-NL3F10H variants as before, however a dialysis step was added in order to remove the excessive Imidazole used for eluting the protein from the Ni-NTA beads. The dialysis was performed at 4ºC over-night. This step was introduced because Imidazole can interfere with the Bradford protein assay, resulting in poor protein quantifications. The elution fractions for the two variants were protein quantified using a linearized version of the Bradford assay (Hayama et al. 2017). This quantification was then used to perform comparable enzymatic assays between the two purified variants.

When making calibration curves, different quantities of known standard are used in order to infer the quantities of the substance of interest in a sample. Therefore, I explored the effect of serial dilutions in the activity of the variants, expecting a proportional drop in activity after each dilution. Surprisingly, I observed lower enzymatic activity in the NL3F10H variant compared to the MBP-NL3F10H in the tested conditions (Figure 5.3).

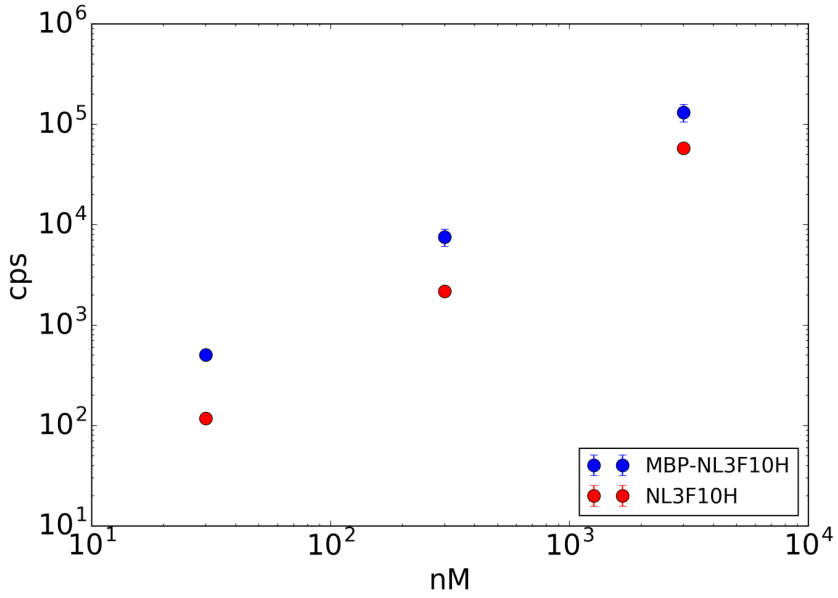


Figure 5.3 Purified NanoLUC versions were quantified by linearized Bradford assay and adjusted to 10mM. The reactions were incubated for 10 min at 21ºC and quantified in a Tristar plate reader with 1.5s signal integration. The purification gel is presented in Figure 5.2. Error bars: S.E.M. counts per second (cps)

This difference could be explained by a possible stabilisation or folding effect by MBP, helping to recover more active NLUC. As this fusion is also more

representative of the TF fusions of interest, it might be better to use MBP-NL3F10H as standard when generating calibration curves for absolute quantification of clock proteins in plant extracts. Also, the MBP adds the advantage of a possible second purification step, using an Amylose column which can be bound by MBP.

I then tested whether plant extracts could inhibit MBP-NL3F10H activity by making serial dilution of MBP-NL3F10H in a crude whole cell extract of 21-day-old Col-0 rosettes. I compared the activity at the $1x10^{-3}$ dilution in order to avoid saturation problems in the plate reader (Figure 5.4). The results show that the signal decreases consistently by an order of magnitude after each dilution in plant extract. Decay in the signal could be appreciated at between the EB and plants extracts at $1x10^{-3}$ dilution.

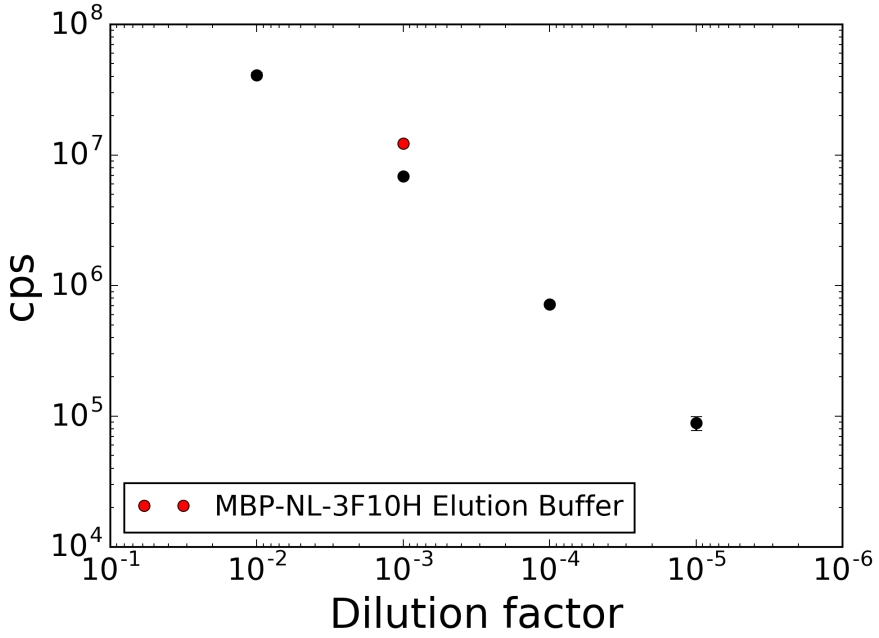


Figure 5.4 Effect of Col-0 plant extract on the activity of MBP-NL3F10H. Purified enzyme by iMAC was diluted in Elution buffer (red) or in plant extract (black). Three technical triplicate assays, mean and error bars SEM.

Working with an enzyme as a calibrator can be problematic if the activity is very unstable. If the enzymatic activity decays in pure preparations, this will result in an overestimation of tagged protein levels in freshly prepared plant extracts. Therefore, I did a simple experiment to test the stability and derive a possible exponential decay model for the enzyme. This model can then be used for correcting quantifications because the enzyme is not used immediately after purification but within 4 days. The results show that MBP-NL3F10H retains more than 95% of activity before three days and more than 80% after one week at 4ºC. A linear regression on the log-transformed time series resulted in a half-life of 37.2 days (Figure 5.5). These results show that

MBP-NL3F10H is a relatively stable enzyme which can be used for generating calibration curves in plant extracts for absolute quantification.

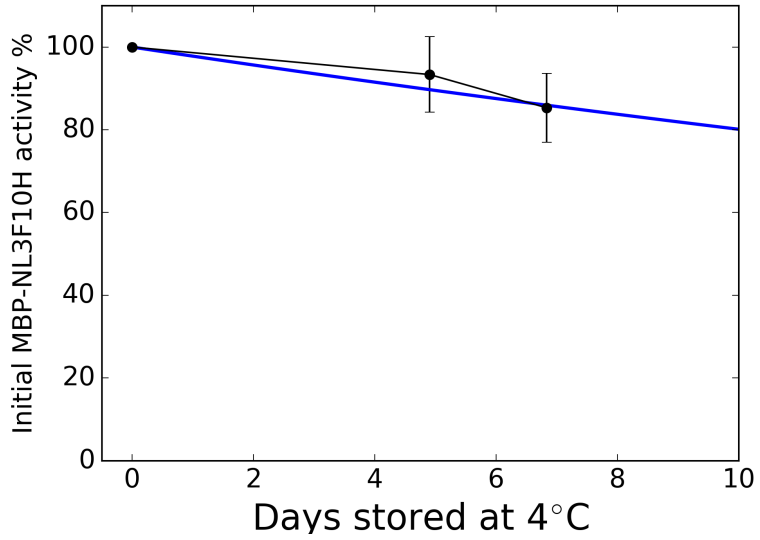


Figure 5.5 Decay of MBP-NL3F10H stored at 4ºC in 250 mM NaCl 50mM NaH_2PO_4 pH 8.0 (NaOH adjusted). The activity was determined by diluting the purified enzyme $1x10^{-3}$ by mixing 20µl of dilution with 80µl of BI buffer and the mixing with 100µl of 1:50 Furimazine:NanoGlow buffer in a 96-well plate. Incubated for 10 min at 21ºC and measured with Tristar plate reader with 1.5s of signal integration. Mean values for three technical replicates errors bars SEM Data are means of error bar shows SEM.

Extending the pGWB vector series for with a NanoLUC3F10H tag.

After an extensive survey in the literature I found the improved pGWB vector series as a very strong candidate starting point for the generation of for tagging proteins with NanoLUC (Nakamura et al. 2010; Nakagawa et al. 2007). In the plant research community, the pGWB vector series is used extensively for transgenic plant production. These vectors are designed for the use of the Gateway recombination cloning technology. The purpose of this work is to test

the functionality of NanoLUC as a reporter, therefore the use of a well-established system eases the interpretation and may reduce trouble shooting during the development of NanoLUC tools. These vectors present several characteristics that make them attractive for testing NanoLUC.

First, I am interested in a vector series for which broad user experience exists. This is a fundamental characteristic because generation of transgenic plants is a very lengthy and time-consuming process. Second, the vector maps are transparent which gives confidence about the elements present in the plasmid. Another important characteristic is that they have a tail to tail construct:marker design, which might help to mitigate unintended effects on the expression of the synthetic construct by the selection marker. Third, the promoter on the selection marker is a NOPALINE SYNTHASE (NOS) promoter rather than a CaMV35S promoter. It has been shown that the CaMV35S promoter can result in secondary effects on the expression of nearby genes, possibly affecting the expression of the construct of interest (Yoo et al. 2005). Fourth, the vectors build on previous work, as extensive libraries of Gateway clones exist. Finally, they present a diversity of selection markers (Kanamycin, Hygromycin, BASTA and Tunicamycin). This is a very important feature as many clock mutants already carry different selection markers, and complementation experiments on these mutants are crucial for testing the functionally of the transcription factors fused to NanoLUC.

I proceeded in the creation of a pGWB with NL3F10H, once the tag was tested and found to be functional. Using pGWB601 as a base vector I introduced NL3F10H by a home-made master mix for Gibson assembly as described in the methods chapter. This resulted in the new pGWB601NL3F10H, with BASTA resistance as a selection marker for plant transformation. The resulting vector allows the generation of C-terminal translational fusions with NL3F10H by Gateway cloning. The nature of the constructs allowed easy sub-cloning into pGWB401, pGWB501 and pGWB701, resulting in a collection of vectors with different plant selection markers (Figure 5.6). The assembly approach and primers used are described in the methods chapter.

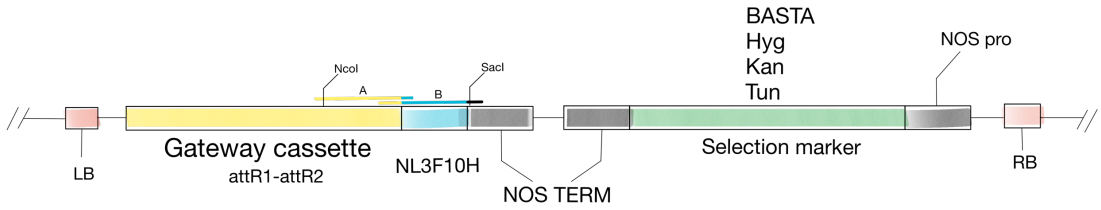


Figure 5.6 Structure of pGWB vectors extended for NL3F10H translational fusions. The Gateway NL3F10H cassette was generated by Gibson Assembly of PCR products A and B, designed with overlapping regions between them and the pGWB601 vector used in the assembly. The vector was digested with Nco/SacI restriction enzymes to insert the new cassette. T-DNA Left Border (LB) and right border (RB). NOPALINE SYNTHASE promoter and terminator (NOS pro, NOS Term). Gateway cassettes of type attR1-attR2 can be used as destination in a LR Gateway cloning reaction. The impGWB vectors have a common Scp^r marker for selection in bacteria, with a pPZP vector backbone. The variants used here were pGW401 (Kan^r), pGWB501 (Hyg^r), pGWB601 (BASTA^r) and pGWB701 (Tun^r). NL3F10H was introduced in each of them by NcoI/SacI sub-cloning.

Testing NanoLUC in *Arabidopsis* protoplasts

In order to test the activity of NanoLUC in plant cells, I cloned the CaMV35S promoter into a pDONR221 vector for subsequent recombination into pGWB601NL3F10 and pGWB635 vectors. The only difference between these two vectors is the tag. The second vector has a firefly luciferase tag (LUC). I used the LUC vector as positive control to rule out problems with the 35S promoter or transformation, if NL3F10H activity is not observed. I purified the vectors containing the 35S:NL3F10H and 35S:LUC constructs, transfected 4-weeks old Col-0 protoplasts and monitored the signal for three days in constant light conditions at 21ºC, in the Tristar plate reader (Figure 5.7).

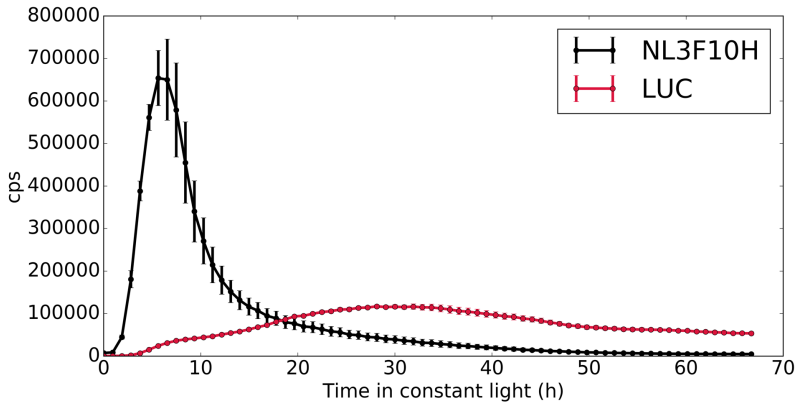


Figure 5.7 Activity of NanoLUC3F10H in protoplasts, compared to firefly luciferase (LUC). Protoplasts were isolated and transformed with equimolar amount of plasmid by the method of (Hansen & van Ooijen 2016). Then followed for three days in constant light at 21C 50µmol/cm²s² monochromatic blue and red LEDs in a 96-well black plate. Furimazine used as substrate for NL3F10H and luciferin for LUC. Mean of three technical replicates error bars show SEM.

These results show that NL3F10H is translated into active enzyme in protoplasts. Interestingly the dynamics of the two constructs differ. It is possible that the strong over-expression of NL3F10H results in substrate depletion compared to the LUC. These constructs can then be used for the generation of transgenic lines overexpressing NL3F10H. Also, notice that signal can be detected much earlier for NL3F10H than for LUC, possibly because of higher brightness of NLUC.

NanoLUC activity in plant extracts of stable transgenic lines

In future experiments NL3F10H will be fused to transcription factors and stable transgenic lines generated for these fusions. Therefore, a protocol for generating plant extracts with active NL3F10H needed to be designed. As a prototyping control, I cloned the strong, viral CaMV 35S promoter into the vector pDONR221 then recombined into pGWB601NL3F10H. I transformed the resulting pGWB601:35S:NL3F10H vector into Col-0 plants as described in methods, selecting with BASTA, resulting in several NanoLUC overexpressing lines. I randomly chose a NanoLUC-positive T3 line for the following optimisation experiments, growing plants for 21 days at 21ºC in a 12L:12D photoperiod of white light at 140μmol/cm^{2}s^2. I tested the activity extracted by manual grinding with a plastic pestle (Figure 5.7A1). I flash froze ??some tissue aliquots?? in liquid nitrogen and stored the rest of the samples at -80ºC. Then after two days in this condition the samples were transferred to liquid nitrogen and extracted using a TissueLyser ball mill, using up to three

grindings for 1 min at 30 Hz, with liquid nitrogen steps between them. The result shows that the extracted activity saturated after the second grinding (Figure 5.7A[2-4]). Also, I tested the scenario where no -80ºC storing took place, with the above extraction approach. Surprisingly, the amount of enzymatic activity recovered is lower after storing the tissue, compared to preparing plant extracts from prepared from plants that were not stored at -80ºC and were processed directly from liquid nitrogen (Figure 5.7B). These results suggest that the samples should be processed on the same day. However more replications should be conducted to in order to properly optimised storage conditions. For the moment in future work I will be conservative and process the samples without storing them at -80ºC for later work.

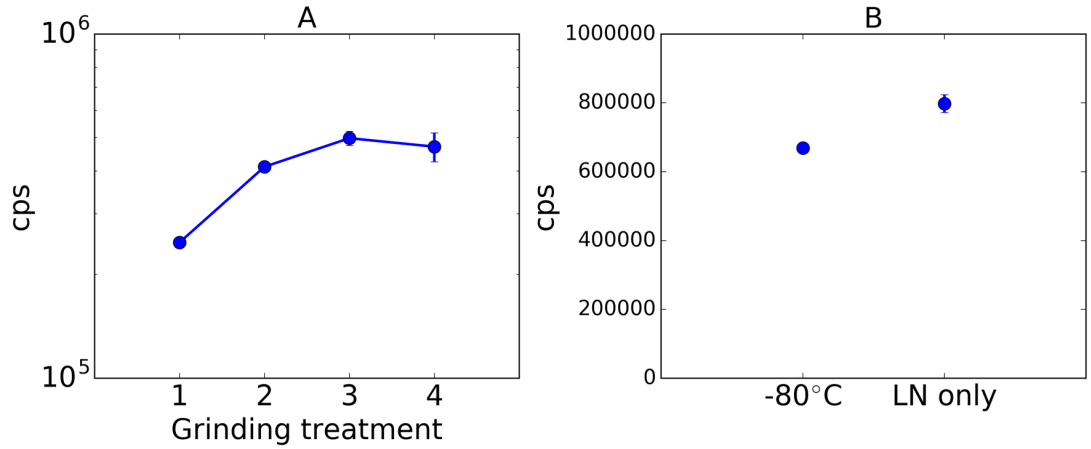


Figure 5.8 Extraction of NL3F10H from tissue of stable transgenic lines. A) Comparison of extraction method, 1) manual grinding, 2) treated once with liquid nitrogen and ground in Tissuelyser. 3) Second round of grinding in tissue lyser. 4) Third round of grinding of tissue lyser. B) Activity decays after storing at -80ºC tissue for two days compared to liquid nitrogen (LN). 10 µl of plant mixed with 190µl of BSII buffer then loaded in a 96-well black plate. Circles represent mean three technical replicated with error bars representing SEM.

NanoLUC data using a poorly characterised clock gene.

A good candidate as an alternative to the LUX protein in the Arabidopsis Evening Complex is its homologue BROTHER OF LUX ARRHYTHMO (BOA). LUC data of BOAp:BOA-LUC indicated that the protein presents a peak of translation at ZT19 (Chapter 3 and 4). Therefore, I created a BOAp:BOA-NL3F10H version to test if the NL dynamics followed the LUC profile. I amplified the BOA promoter and coding region as described in the method section, recombined the pDONR221:BOAp:BOA with pGWB601NL3F10H and tested for NLUC activity in Col-0 protoplasts after verifying the activity of NL in the transient system. I transformed fusions of BOAp:BOA-NL into Col-0 plants carrying CCA1p:LUC, in order to follow the effect of BOA on a well-established clock reporter gene, and transformed BOA-LUC into Col-0 wild-type plants. I performed Agrobacterium mediated transformation by floral dipping and obtained transgenic T3 homozygous lines as described in the methods Chapter. Using homozygous T3 lines I conducted a plate reader experiment for *in vivo* tracking. This approach is widely used for LUC clock reporters, where luciferin can be applied to plants in a solution containing 0.01% Triton X-100.

I inoculated seeds in a black, 96-well plate and stratified them for 2 days at 4ºC. The plate was then transferred to 12L:12D photoperiod in 140 μmol/cm^{2}s^2 white light at 21ºC in order to obtain plants with short hypocotyls. After 2 days

in 12L:12D photoperiod at 50 μmol/cm^{2}s^2 monochromatic blue and red LEDs at 21°C, so the plants adapt to the measuring conditions, I started automatic signal recording in a plate reader with a 30-min interval between measurements (Figure 4.9). After one diurnal 12L:12D cycle the light settings were changed to constant light and bioluminescence signal was recorded for another 5 days. The genotypes under investigation are two lines of Col-0 CCA1p:LUC BOAp:BOA-NL3F10H, two lines of Col-0 BOAp:BOA-LUC and Col-0 CCA1p:LUC (as control) (Figure 4.9).

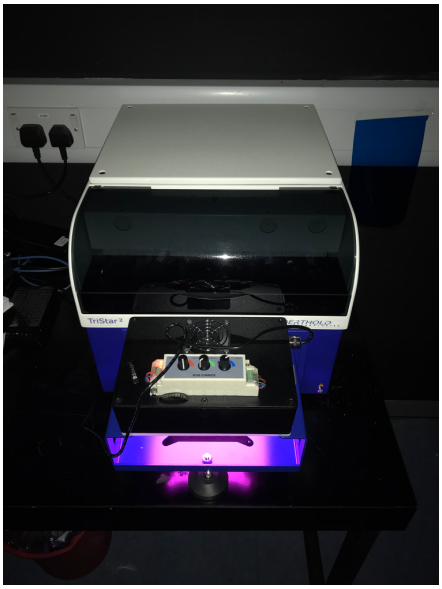


Figure 4.9 Plate reader set up for circadian time series. In between measurements a 96-wells plate sits under a screen of blue and red LEDs in an ambient temperature-controlled room. The intensity conditions are 17 µmol/cm^{2}s^2 blue and 29 µmol/cm^{2}s^2 red light.

This experiment showed, first, that signals can be recorded from the NL3F10H fusion. This is very important as it provides evidence for using *in vivo* measurements of protein levels. BOA fusions to either NanoLUC or firefly LUC

gave similar signal levels, an order of magnitude lower than the CCA1p:LUC transcriptional reporter. Second, BOA-NL3H10 does not present a strong rhythm in LD (Figure 5.11 A). The NanoLUC signal falls in darkness, contrasting with the small increase in firefly LUC signal at ~ZT17h (Figure 5.10 B). Nonetheless an apparent peak of BOA-NL3H10 signal can be observed in antiphase to CCA1p:LUC profile in LL, which is more evident on the 2nd and 3rd day in LL, e.g. at ~15h. This contrasts with the clear rhythmic changes of BOAp:BOA-LUC (Figure F.12B). Another interesting feature is that BOA-NL3H10 expression has an effect on CCA1p:LUC, which increases its rhythmic period. These preliminary data suggest that BOA-NL is an active translational fusion. In the diurnal phase I observe a small drop in signal for both independent BOA-NL lines at lights-off, reversed at lights-on, but it is unclear if it is a biological response of BOA or an artefact in switching off the lights (Figure 5.10A). A similar very subtle effect is observed for BOA-LUC (Figure 5.10B).

This data is surprising as it seems that BOA might present similar dynamics to LUX, which has also been observed to present low amplitude oscillations (Nusinow et al. 2011). An important difference is that LUXp:LUC (PCL1:LUC) presented a peak of expression at ZT12 (Onai & Ishiura 2005), similar to the LUX RNA. This contrasts with BOAp:BOA-LUC, that peaks at ~ZT19 in LD (Figure 5.11).

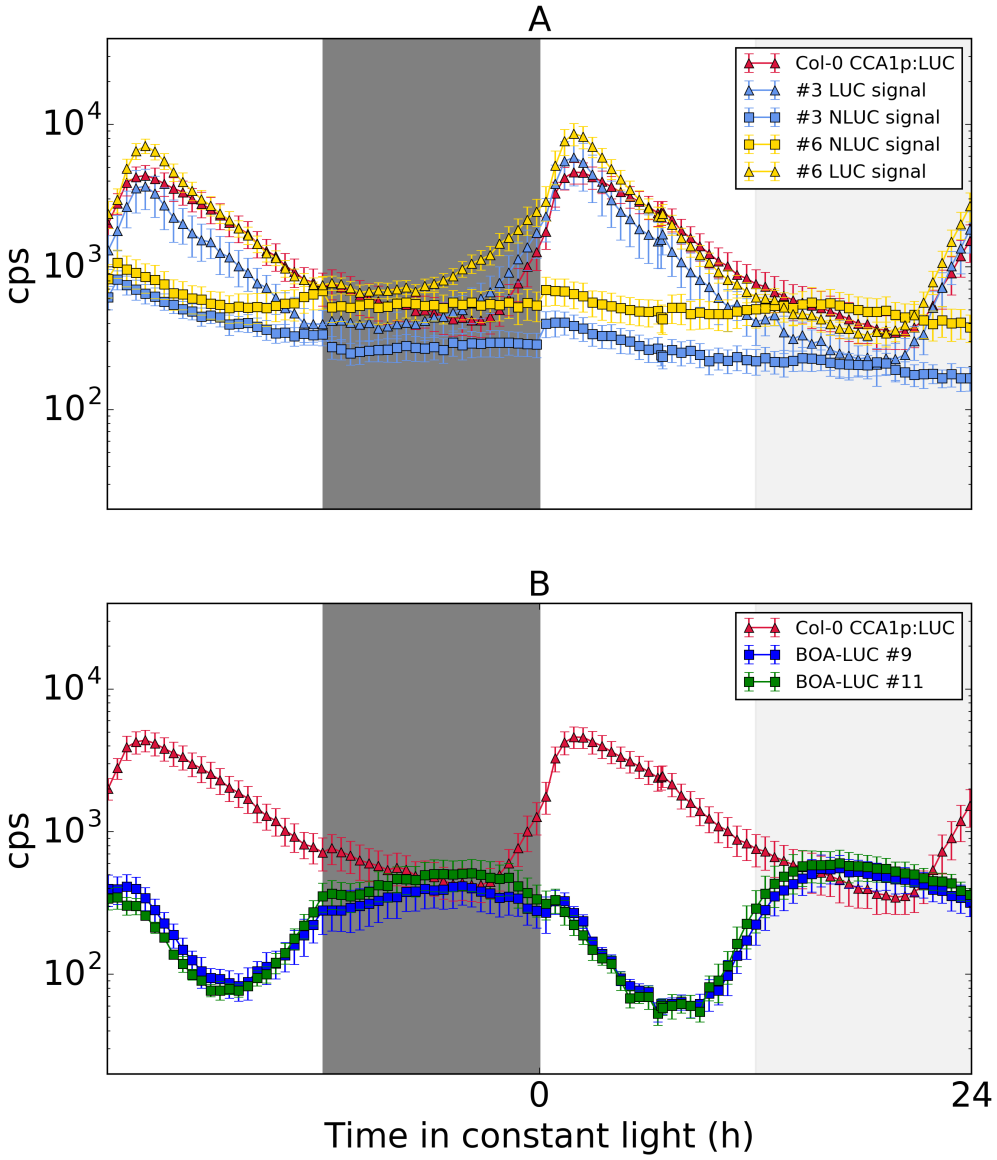


Figure 5.10 Diurnal dynamics of NanoLUC and firefly LUC fusions to BOA. A) Firefly LUC signal from CCA1p:LUC (triangles) in the parent line (red) and two BOA-NL transformants (blue, yellow), compared to NanoLUC signal (squares) in replicate plants of the same BOA-NL lines. A drop in BOA-NL can be observed in darkness, possibly also observed in LUC lines. This might be an artefact of having lights on while measuring in in constant light. 10 seedlings per trace entrained for 7 days at 140 µmol/cm²s² white light 12L:12D 21ºC. Then transferred for 2 days into 50 µmol/cm²s² monochromatic blue and red LEDs. Bioluminescence quantified in a Tristar every 30 min and 1.5s of signal integration. Mean of four assays error bars SEM

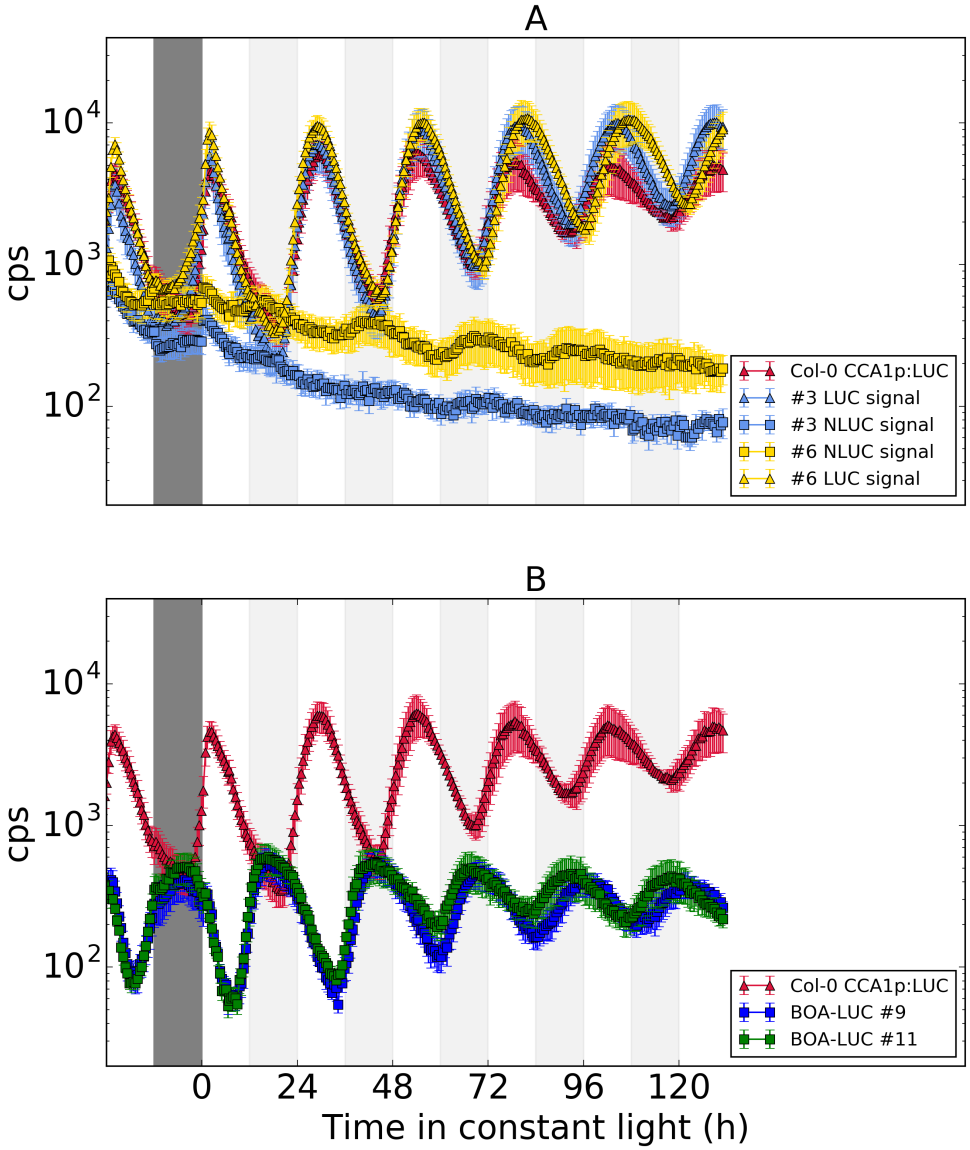


Figure 5.11 Circadian rhythms of BOA protein using NanoLUC and LUC fusions. A) Dynamics of NanoLUC and LUC signal in Col-0 CCA1p:LUC BOAp:BOA-NL3F10H. Two independent T3 lines. B) Dynamics of BOAp:BOA-LUC in Col-0. 10 seedlings per trace entrained for 7 days at 140 µmol/cm²s² white light 12L:12D 21ºC. Then transferred for 2 days into 50 µmol/cm²s² monochromatic blue and red LEDs. Bioluminescence quantified in a Tristar every 30 min and 1.5s of signal integration. The traces are means of 4 replicates assays error bars represent SEM.

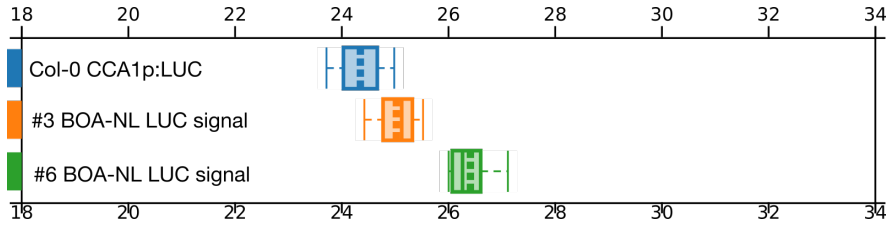


Figure 5.12 BOA-NL impacts the period of CCA1p:LUC. Period of time series in (Figure 4.4). Period determined in Biodare2 using FFT-NLS method. The period was measured from data after the second day in constant light conditions.

In the transformation process, random genomic insertions result in different levels of BOA expression. This can be observed in the mean signal level of the BOA-NL lines and in an effect on the period of CCA1p:LUC (Figure 5.13). The two lines tested provide only preliminary data but the levels of expression of BOA increased the period, which is consistent with previous results reported for CaMV35S:BOA (Dai et al. 2011). The lack of oscillations in BOA-NL3F10H is concerning, as it is possible that the tag affected the degradation rate of the protein. Dai et al reported western blot data at ZT27 (trough) and ZT39 (peak) but very low levels were observed at both time points.

Discussion

The use of tagged lines with general epitopes e.g. GFP, FLAG 10xHis has been a powerful approach to study the protein component of the Arabidopsis

circadian oscillator. Thanks to these lines, several aspects of clock protein regulation have been investigated, including localisation, binding sites in the genome by ChIP, and protein-protein interactions by co-IP or more recently affinity purification mass spectrometry e.g. the case of ELF3 and GI. (Huang et al. 2016; Krahmer et al. 2017). The type of modelling intended in this work requires data in absolute numbers. Quantitative Western blots could have been used for collecting this type of data. However, this approach is laborious and almost impractical for obtaining quantifications in long circadian time series. Therefore, I explored the development the use of the new luciferase NanoLUC (NL) for performing absolute quantifications by assaying the activity of NL fused transcription factors of interests. This approach is quantitative and could potentially provide a way to automatize the generation of protein data with quality similar to qRT-PCR.

In this chapter, I presented the development of synthetic biology NL tools for tagging and performing in future work absolute quantification of clock transcription factors from plant extracts by a reference curve. Several elements were required for establishing these tools in plants. First, I showed that the sequence provided by Promega produces active enzyme in bacteria, protoplasts and stably transformed Arabidopsis lines. It is important to note that for this approach the NL enzyme can be purified to homogeneity in a straight forward way. I achieved this by setting up a straightforward and standard protocol based on Ion Metal Affinity Chromatography (IMAC) (see

methods chapter). The approach can yield a homogenous preparation of NanoLUC N-termini fused to Maltose Binding Protein and C-termini fused to 3xFLAG 10XHis (MBP-NL3F10H). This reagent can then be used for creating calibrating curves using known amounts of NL. The resulting curve can be used for inferring the concentration of transcription factor fused to NL in plant extracts. However, it is important to remark that the method used for determining quantity of purified NL can be error prone and protein amount in the plant extract could be over or underestimated. Therefore, determination of NL-standard concentration needs to be cross-validated with other methodologies in future work, for example by quantitative mass spectrometry or radiolabelling. This should result in more confident quantifications.

The decay rate of NL in the storage conditions tested (37 days) was significantly lower compared to the previously reported 7.7 days at 37ºC (Hall et al. 2012). However, the decay observed at 4ºC might be enough to be considered relevant when inferring the absolute copy number of transcription factors from NL calibration curves. Reduction in enzyme activity therefore requires that the purification takes place on the day the quantification will be performed, a challenging step for a single user when performing quantifications for circadian time series with ideal sampling frequencies (every 2 hours). However, a model of exponential decay can be used for correcting the absolute quantifications, although careful logging of sample preparation times is required when using this approach.

An important issue regarding the transformation of the translational fusion constructs is the need of mutant backgrounds where the tagged protein is absent. In Arabidopsis these mutant backgrounds were generated by T-DNA mutagenesis Salk lines or other methods that exploit the use of an antibiotic marker. These mutant backgrounds in many cases in Arabidopsis were generated by the insertion of a selectable marker, e.g. the T-DNA Salk lines that present resistance to Kanamycin. In some scenarios using double mutants for the clock mechanisms for studying the translational fusion activity by genetic complementation experiments. Furthermore, in order to study the perturbation on the clock by the new construct a LUC gene fused to the promoter of a clock gene or clock output is used. This could result in the need of a fourth selection marker for the translational fusion. In order to avoid such I extended the pGWB vector series adding NanoLUC 3xFlag 10xHis (NL3F10H) as tag in these collection of destination vectors, allowing the use of four different antibiotic selection markers.

Furthermore, many plant labs have collections of Gateway entry clones that can be easily recombined into these new NL3F10H vectors using the Gateway technology. Thanks to the wide use of the pGWB vectors, I could focus on the NL activity rather than vector construction. However, these new vectors can also serve as a reference point if NL is later used in a different plant synthetic biology context, for example when using other assembling methodologies (Golden Braid or Gibson Assembly). Nonetheless, the big issue with this

approach is that the creation of N-termini translational fusions is more complex, which might be required if the C-termini tagging compromises the activity of the protein of interests.

Another problematic scenario using a tagging strategy is the presence of alternative splicing in eukaryotes. The presence of different splice forms for several clock genes e.g. LHY, CCA1 and other core clock genes have been shown to temperature dependent (James et al. 2012). This complicates the generation of fusions as some splicing forms might leave the tag out therefore only reporting a subpool of the total protein. Another problem associated with the strategy is the use of a standard 3'-UTR (NOS terminator). It is known that the 3'-UTR can present important regulatory sequence for example microRNA target sequences that can modulate RNA stability or translatability, therefore modification of them could potentially result in altered transcript and protein dynamics (Jones-Rhoades et al. 2006). In the next chapter I show that it was possible to obtain absolute quantifications for the key clock genes selected. Nonetheless only in combination with a careful theoretical and numerical argumentation confidence on the experimental results could be increased.

This contrast with the example presented in this chapter for BOA. The biological richness of protein synthesis complicates data interpretation of BOA-LUC and NL fusions which present different. The LUC version present clear oscillations while the NL fusion show barely detectable rhythms. The LUC data

suggest that that the transcript encoding LUC is rhythmic. However, it is possible that the BOA-LUC signal presents strong rhythmic behaviour given that LUC half-life is short. This can be observed in the clock promoters fused to LUC e.g. CCA1p:LUC. Further experiments need to be performed to clarify if the lack of oscillation is a real feature of BOA or an experimental artefact generated by the construct. Alternatively, it is possible that the particular lines selected result in misexpression of the transcript at constant levels. This can be verified by performing qPCR for the transgenic construct and comparing it with the endogenous BOA gene. Also, it is possible that the NL tagged is excised by a posttranslational process which can be verified by Western blots.

Another possibility is that BOA fusions present different decay constants compared to WT, NL having a lower one. If this is combined with a small enough translation rate, then low amplitude oscillation can be obtained. It is possible to observe that BOA-LUC tagged lines have much lower signal levels compared to CCA1p:LUC. This indicates that it is possible that BOA-LUC has very low translation rate, which is consistent with a scenario where low translation and low degradation results in low amplitude. It could be possible to test this by 35S-methionine radiolabelling of the different BOA proteins with an in-vitro transcription and translation system like PURE offered commercially by New England Biolabs. The synthesised protein variants can then be incubated in Col-0 plant extracts at the different times of day and follow the decay kinetics using autoradiography Western blots, comparing them will allow

testing if the decay constant has been affected by the introduction of the C-termini tag. Also, a N-termini tagged version can be tested as an alternative version in case C-termini tagging affects BOA protein stability. Despite the low amplitude of BOA-NL, LUC dynamics suggest that the other construct elements (promoter and terminator) respected the phase of expression of transcript. This suggests that BOA might contribute late at night as an element of the Evening Complex and therefore complementary to LUX.

The previous discussion exemplifies the issues associated with using tagged lines for reporting protein dynamics compared to the more stablished transcriptional reporters which in principle should present a less pronounced impact on the dynamics of the system. Even though at the moment several points need to be clarified regarding the dynamics of the BOA protein the results shows a genetically active construct affecting the behaviour of CCA1p:LUC reporter. This is consistent with previous reports using BOA overexpressing lines (Dai et al. 2011), indicating that the fusion protein is transcriptionally active. Also, this result shows that bioluminescence assays with different substrates can follow either the period of the transcriptional reporter or the levels of the transcription factor, among plants of a single transgenic line. This could allow in principle the generation of an experimental sensitivity analysis where the impact of protein levels on the period of the oscillator could be studied, as it has been done for ZTL using western blots (Somers et al. 2004). The reporter fusion proteins offer a unique opportunity

to link absolute quantities of transcription factors and their impact on interesting physiological features, including output phenotypes that can be used for example in the multiscale models where the clock has been incorporated (Chew et al. 2014; Chew et al. 2017).

References

Chew, Y.H. et al., 2017. Linking circadian time to growth rate quantitatively via carbon metabolism. *bioRxiv*, p.105437.

Chew, Y.H. et al., 2014. Multiscale digital Arabidopsis predicts individual organ and whole-organism growth. *Proceedings of the National Academy of Sciences of the United States of America*, 111(39), pp.E4127–36.

Corellou, F. et al., 2009. Clocks in the green lineage: comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. *The Plant cell*, 21(11), pp.3436–3449.

Dai, S. et al., 2011. BROTHER OF LUX ARRHYTHMO is a component of the Arabidopsis circadian clock. *The Plant cell*, 23(3), pp.961–972.

Feeney, K.A. et al., 2016. Daily magnesium fluxes regulate cellular timekeeping and energy balance. *Nature*, 532(7599), pp.375–379.

Fogelmark, K. & Troein, C., 2014. Rethinking transcriptional activation in the Arabidopsis circadian clock. *PLoS computational biology*, 10(7), p.e1003705.

Fujiwara, S. et al., 2008. Post-translational regulation of the Arabidopsis circadian clock through selective proteolysis and phosphorylation of pseudo-response regulator proteins. *The Journal of biological chemistry*, 283(34), pp.23073–23083.

Hall, M.P. et al., 2012. Engineered luciferase reporter from a deep sea shrimp utilizing a novel imidazopyrazinone substrate. *ACS chemical biology*, 7(11), pp.1848–1857.

Hansen, L.L. & van Ooijen, G., 2016. Rapid Analysis of Circadian Phenotypes in Arabidopsis Protoplasts Transfected with a Luminescent Clock Reporter. *JoVE (Journal of Visualized Experiments)*, (115), pp.e54586–e54586.

Hayama, R. et al., 2017. PSEUDO RESPONSE REGULATORs stabilize CONSTANS protein to promote flowering in response to day length. *The EMBO journal*, 36(7), pp.904–918.

Herrero, E. et al., 2012. EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the Arabidopsis circadian clock. *The Plant cell*, 24(2), pp.428–443.

Huang, H. et al., 2016. Identification of Evening Complex Associated Proteins in Arabidopsis by Affinity Purification and Mass Spectrometry. *Molecular & cellular proteomics : MCP*, 15(1), pp.201–217.

James, A.B. et al., 2012. Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *The Plant cell*, 24(3), pp.961–981.

Jones-Rhoades, M.W., Bartel, D.P. & Bartel, B., 2006. MicroRNAS and their regulatory roles in plants. *Annual review of plant biology*, 57, pp.19–53.

Krahmer, J. et al., 2017. Time-resolved Interaction Proteomics of the Putative Scaffold Protein GIGANTEA in Arabidopsis thaliana. *bioRxiv*, p.162271.

Mas, P. et al., 2003. Targeted degradation of TOC1 by ZTL modulates circadian function in Arabidopsis thaliana. *Nature*, 426(6966), pp.567–570.

Millar, A. et al., 1992. A novel circadian phenotype based on firefly luciferase expression in transgenic plants. *The Plant cell*, 4(9), pp.1075–1087.

Nagel, D.H. et al., 2015. Genome-wide identification of CCA1 targets uncovers an expanded clock network in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34), pp.E4802–10.

Nakagawa, T. et al., 2007. Improved Gateway binary vectors: high-performance vectors for creation of fusion constructs in transgenic analysis of plants. *Bioscience, biotechnology, and biochemistry*, 71(8), pp.2095–2100.

Nakamura, S. et al., 2010. Gateway binary vectors with the bialaphos resistance gene, bar, as a selection marker for plant transformation. *Bioscience, biotechnology, and biochemistry*, 74(6), pp.1315–1319.

Nusinow, D. et al., 2011. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

Onai, K. & Ishiura, M., 2005. PHYTOCLOCK 1 encoding a novel GARP protein essential for the Arabidopsis circadian clock. *Genes to cells : devoted to molecular & cellular mechanisms*, 10(10), pp.963–972.

Puigbò, P., Bravo, I.G. & Garcia-Vallvé, S., 2008. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC bioinformatics*, 9(1), p.65.

Shen, J. et al., 2013. Organelle pH in the Arabidopsis endomembrane system. *Molecular plant*, 6(5), pp.1419–1437.

Somers, D.E., Kim, W.-Y. & Geng, R., 2004. The F-box protein ZEITLUPE confers dosage-dependent control on the circadian clock, photomorphogenesis, and flowering time. *The Plant cell*, 16(3), pp.769–782.

Yoo, S.Y. et al., 2005. The 35S promoter used in a selectable marker gene of a plant transformation vector affects the expression of the transgene. *Planta*, 221(4), pp.523–530.

Chapter 6
Absolute quantification of Arabidopsis Clock proteins

"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong" – Richard P. Feynman-

In chapter 4, I performed a theoretical exercise to understand how absolute units for the protein components could be introduced in clock models, using protein levels that were predicted from RNA data in absolute units. This approach uncovered unrealistic assumptions that had been made in the models leading up to U2017.2 and showed that CCA1, PRR7 and LUX protein measurements were the minimum required to constrain the mass scale. In the previous chapter, I used equilibrium binding data to test if the predicted protein levels were in a reasonable range compared to the DNA-binding affinities of the proteins. Integrating Chip-seq, PBM and SPR data suggested that the predicted levels were consistent with observed gene regulation. Nonetheless it remained important to challenge the predicted levels by gathering temporal data on the absolute number of clock protein molecules per cell.

Quantitative mass spec proteomics has more recently emerged as a means to quantify clock proteins, though this presents important technical challenges (Narumi et al. 2016). Preliminary work with the Gruissem lab had failed to detect clock proteins by a sensitive SRM methodology, even in over-expressing transgenic plant lines (Graf and Gruissem, TiMet project, unpublished results), so this possibility was not pursued here).

Instead, I opted for a more classical approach for gathering protein data in absolute units, with quality to qRT-PCR measurements. The proteins of interest are fused to a reporter, for which a calibration curve with known amounts of a standard can be used to infer the amount of tagged protein. Previous work with LUC fusions has been important for obtaining waveforms rather than absolute levels e.g. in mouse Per2::LUC(Yoo et al. 2004), in plant PIF1:LUC (Shen et al. 2005), and later in algal CCA1 and TOC1 (Corellou et al. 2009). Reporter fusions avoid the need for specific antibodies for each protein (which were used to quantify mammalian clock proteins, (Lee et al. 2001), but have proved more challenging in plants, notably for TOC1), allow the selection of a sensitive and tractable reporter assay, however requires time-consuming transgenic plant production, is indirect (the clock protein could be cleaved from the reporter) and does not test native protein in any plant of interest. The transgene delays testing in the multiple mutants that have been critical in understanding plant clock function. Anticipating this concern, I used multiple-mutant backgrounds that will speed up later generation of a range of

mutant lines containing the transgene. Also, using these lines is an important step for testing the functionality of the transcription factor by complementation experiments.

Therefore, I exploited the use of the NanoLUC (NL) for performing absolute quantification using the tools developed in Chapter 3. I generated time series for the key clock genes proposed in Chapter 4. I present the implications of the quantifications I obtain for the U2017.1 model. Using homozygous lines for of NL fusions I generated 72h time-series that present both diurnal and circadian constant light regime. I decided to capture the transition from entrainment into constant conditions as it has been shown to be an very informative sampling strategy to infer model parameters from time series (Mombaerts et al. 2016). Also it allows comparison with TiMet RNA data (Flis et al. 2015). This work resulted in a gold standard time series in absolute units for the proteins of interests. The resulting quantification allowed me to test the predicted levels of proteins from Chapter 3. The implications of these quantifications for the clock model are treated in the result section. Overall the method developed in constitute a mayor advance for challenging plant clock models with high quality protein data in absolute units.

Results

Design of translational fusions

First, I created the required constructs which are fusions NanoLUC 3xFlag 10xHis to the C-termini of LHY, CCA1, PRR7, TOC1 and LUX (Figure 5.1). The constructs are genomic fragments that include the cognate promoters for the genes of the latter proteins. I decided to retain intronic regions because it has been observed that extensive alternative splicing takes place in a temperature dependent way (James et al. 2012). For example, for CCA1 two iso-forms exist α and β. The β iso-form lacks the MYB DNA binding domain, and inhibits the activity of the alpha isoform (Seo et al. 2012). From the multiple isoforms, I selected those that produce the longest coding site.

Single clock mutants for CCA1, LHY and PRR7 have mild period therefore I decided to use double mutants as a first approach to test the transcription factor activity of the NL fusions. Once the fusions are tested they can be crossed to single mutants to generate the correct genotypes in future work. Experimentally speaking there are several difficulties associated with the generation of these new transgenics. For example, construct will require a collection of homozygous lines in order to find complementing candidates. This is a consequence of the random insertion during the transformation process with *Agrobacterium tumefasciens* (van Kregten et al. 2016). Therefore, more than a dozen lines are required for finding a period rescuing line. The randomness of the integration process can be seen as a drawback.

Nonetheless this process results in different expression levels which can be exploited in an experimental sensitivity analysis for the transcriptional parameters in the Arabidopsis oscillator.

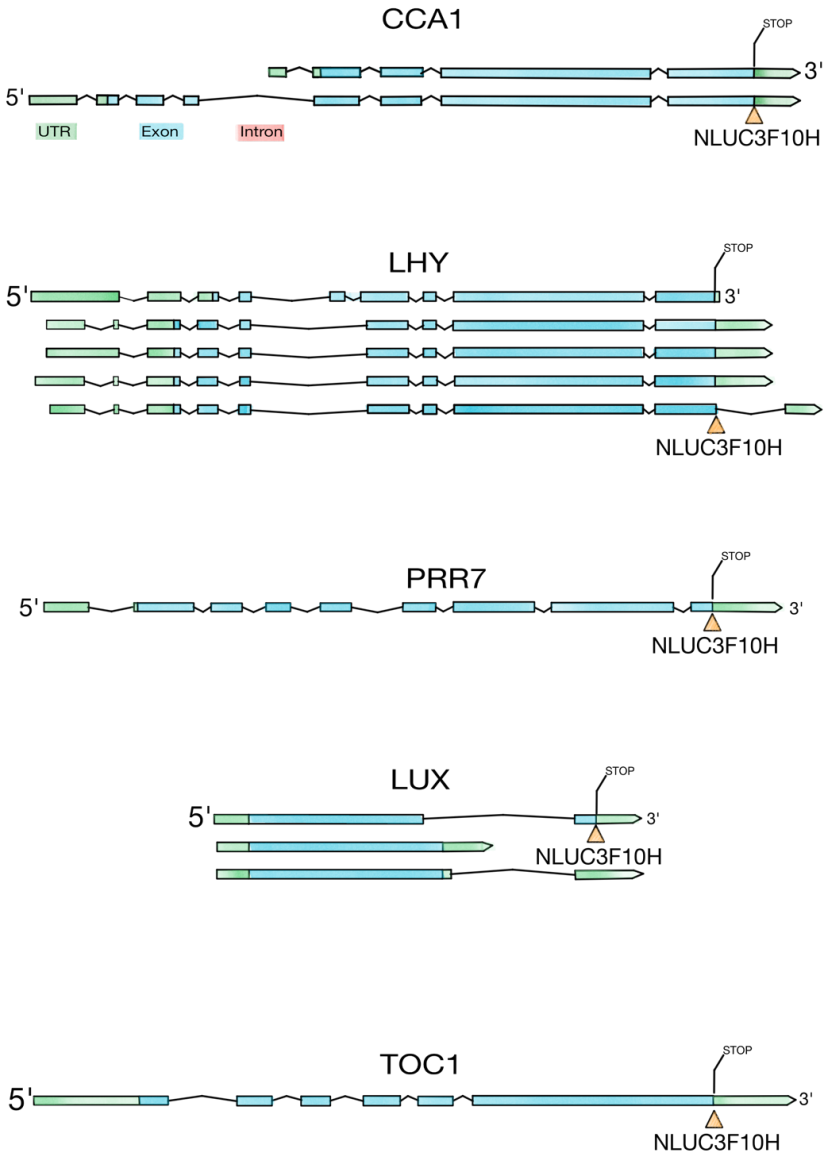


Figure 6.1. Genomic structure, selected mRNA isoforms and location of NanoLUC insertion for the clock genes of interest. The orange triangle indicates which transcript model was used as reference for the fusions. The stop codon was removed and replaced with NanoLUC3F10H. The terminator regions were replaced by a common NOS-T region.

Prototyping synthetic constructs in protoplasts

I conducted preliminary experiments using a protoplast transient system before engaging in the lengthy process of transgenic plant generation. For this propose I transformed Col-0 protoplasts using the method (Hansen & van Ooijen 2016). As a positive control I also transformed CaMV35S:NL3F10H and CaMV35S:LUC. In addition to the NanoLUC3F10H fusion, LUC fusions were tested for comparative purposes, and all the constructs showed strong light emission in protoplasts compared to the empty vectors (Table 1). This data suggested that the construct architecture did not disrupt the capacity for the gene to be transcribed and translated. However, I did not observed oscillations in protoplast experiments using Col-0 protoplasts. Nonetheless, I did not pursue the what could be the underlying problem as the main focus is generating stable transgenic, which can be investigated in future work. To the best of my knowledge there is now report of period complementation in a protoplast system. I proceeded with generation of stable transgenic lines through rescuing clock mutants (7-12 months of work).

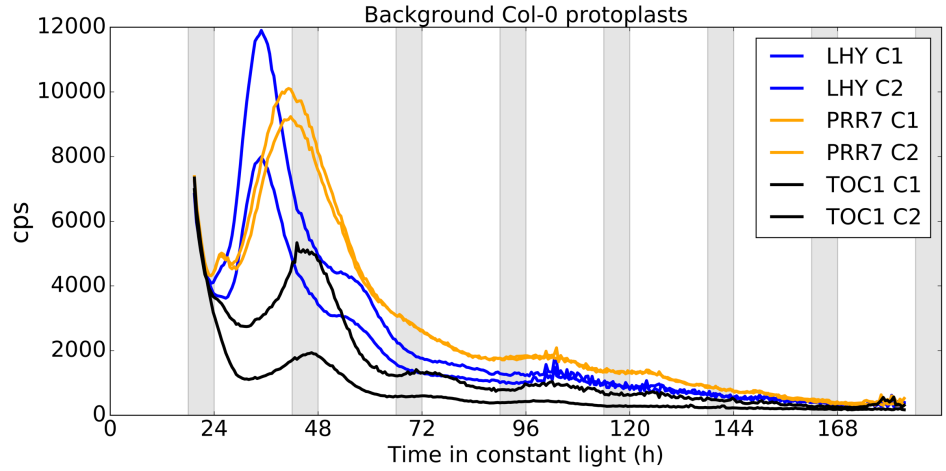


Figure 6.2 Transient expression in protoplast of several constructs of interest. Two independent clones from the first PCR were tested to verify reproducibility of results. Correct order of expression can be observed at 24-48h: LHY,PRR7 and TOC1 peaked in that order. The photoperiod was 16L:12D for speeding up the generation of protoplasts.

Table 6.1 Constructs tested in Col-0 protoplasts by the method of Hansen et al.

Vector	Reporter	Signal
pGWB601::35S:NanoLUC3F10H	NanoLUC	Yes
pGWB635::35S:LUC+	LUC+	Yes
pGWB601::LUXproLUX-NL3F10H	NanoLUC	Yes
pCambia1305.1::CCA1proCCA1-NL3F10H	NanoLUC	Yes
pGWB635:LUXproLUX-LUC+	LUC+	-
pGWB601::CCA1proCCA1-NL3F10H	NanoLUC	Yes
pGWB601::TOC1proTOC1-NL3F10H	NanoLUC	Yes
pGWB635::CCA1proCCA1-LUC+	LUC+	Yes
pGWB601::LHYproLHY-NL3F10H	NanoLUC	Yes
pGWB601::PRR7proPRR7-NL3F10H	NanoLUC	Yes
pGWB635::TOC1proTOC1-LUC+	LUC+	-
pGWB635::PRR7proPRR7-LUC+	LUC+	Yes

Testing for complementation using transgenic

I transformed performed the following plant transformations by floral dipping. CCA1p:CCA1-NL and LHYp:LHY-NL -> lhy-1/cca1-11 (Kamioka et al. 2016), PRR7p:PRR7-NL -> *prr9/prr7* (Farré et al. 2005) TOC1p:TOC1-NL *toc1-2* (Strayer et al. 2000) LUXp:LUX -> *lux-2* (Hazen et al. 2005). A collection of homozygous lines is essential to select candidates for complementing the clock. These were obtained by segregation analysis using BASTA or Bialaphos (see Methods chapter). With homozygous lines, I performed imaging experiments to determine the period in constant light conditions, for the NanoLUC fusion lines as described in the methods section (Figure 5.1). The effect of the constructs on the period of clock mutants can be observed compared to the mutant parents. CCA1-NL and LHY-NL were able to increase the period of lhy-1/cca1-11 (Figure 5.5A). The PRR7-NL construct was able to shorten the period phenotype of the *prr7,prr9* mutant (Figure 5.5B). But it is important to notice that the period of the PRR7-NL tend to accumulate above 25h, with some exceptions. *toc1-2* presents a short period phenotype and TOC1-NL increased the period (Figure 5.5C). Finally, the arrhythmic phenotype of *lux-2* is rescued by the LUX-NL constructs (Figure 5.5D). From this collection of lines several were selected for generating time series of absolutes quantifications (Table 5.1). Thanks to this selection approach I was able to obtain enough diversity in the transgene expression of each construct for finding goof candidates for absolute quantifications.

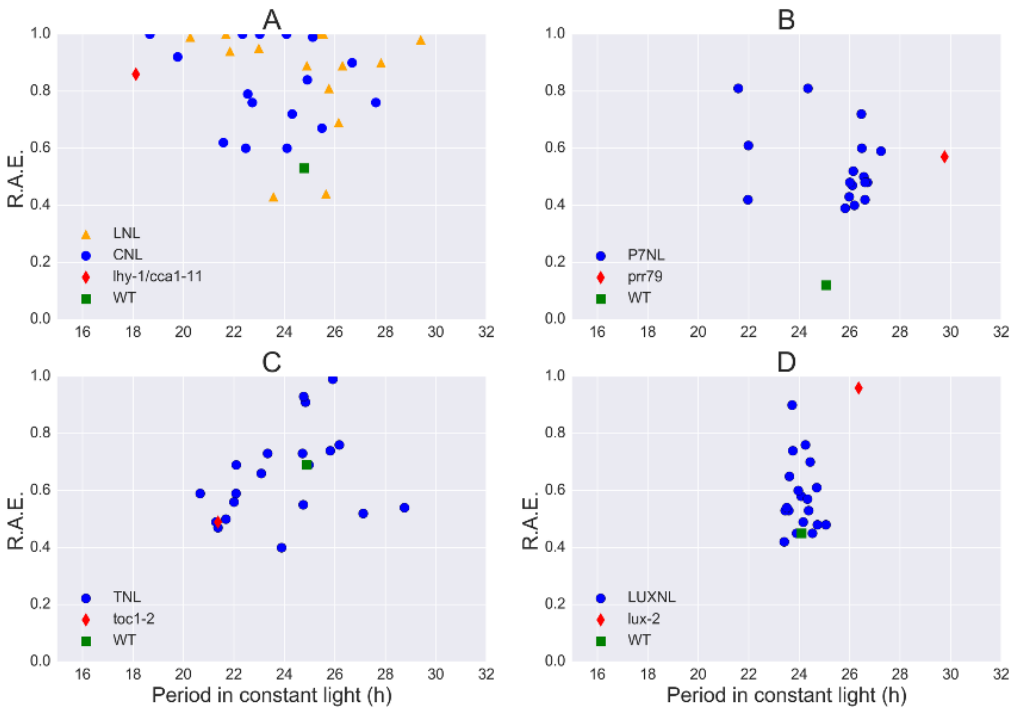


Figure 6.3 Period analysis of T3 homozygous NanoLUC lines. A) LHY and CCA1 NanoLUC. B) PRR7 NanoLUC C) TOC1 NanoLUC and D) LUX NanoLUC. The periods of the lines used for generating the time series are summarised in Table 6.2

Table 6.2 Summary of periods for lines used for absolute quantification, the period controls for each of the fusions are also indicated WT and mut above selected lines.

Genotype	*Period*	*Period Std*	*ERR*
Col-0 *CCA1p:LUC WT*	24.78	1.6	0.53
delt cl *CCA1p:LUC mut*	18.11	0.44	0.86
delta cl *CCA1p:LUC LNL35.1*	25.64	0.14	0.44
delta cl *CCA1p:LUC CNL63.1*	25.48	1.82	0.67
Col-0 *CCR2p:LUC WT*	24.88	1.08	0.69
Col *toc1-2 CCR2p:LUC mut*	21.36	0.46	0.49
Col *toc1-2 CCR2p:LUC TN30.2*	24.96	0.03	0.69
Col-0 *CCR2p:LUC P7NL18.2*	25.82	1.08	0.39
Col-0 *CAB2p:LUC WT*	24.37	0.07	0.53
Col *lux-2 mut*	26.35	2.76	0.96
Col *CAB2p:LUC LUXNL21.2*	23.58	0.44	0.53
Col *CAB2p:LUC LUXNL9.2*	24.52	0.23	0.45

Circadian time-series in absolute units for clock proteins.

After the intensive work of generating new transgenic lines I selected a series of candidates to determine absolute amount of total protein per cell using NanoLUC signal. I stratified seed for 48 hours and plated individual seeds on ROBUST media. Plants were transferred at ZT0 to a light regime of 12L:12D at 21ºC and ~ 110 μmols/cm^{2}s^2. After two weeks plants were transplanted to soil and grown for another week in the same light conditions. Samples were harvested every two hours starting 15 min before lights-off (ZT12) for 24 hours. Light conditions were changed to constant light for the rest of the experiment with sampling taking place every two hours. Two biological replicates were taken, each consisting of 5 pooled plants, to produce enough material for plant extracts. The harvested tissue was maintained in liquid nitrogen until processed for determination of NanoLUC activity as described in the methods chapter. I inferred the number of molecules per cell by generating a calibration curve of purified MBP-NL3F10H protein as described in the methods Chapter. Resulting in the first time-series of core clock protein with absolute number of molecules per cell (Figure 5.11).

These results are encouraging as all the protein signals are stably rhythmic, compared to the damped rhythms in protoplast for LHY, PRR7 and TOC1 proteins (Figure 5.8). The proteins show peaks of expression similar to previous reports. The total number of samples in the experiment are 360, 72

samples per transgenic line. Unfortunately, I lost 11 samples during tissue grinding the Tissue Lyser (Qiagen). The grinding strength was enough for weakening the cap wall, therefore allowing air to penetrate the damaged tubes. This can be avoided by rapidly opening the tubes after grinding the tissue. However, in general the quality of the data provides was not compromised. I managed to control de variability in the experiment and clear oscillations are observed. Higher variability was observed for the biological replicates than for the technical replicates for when I determined NanoLUC activity in plant extracts. The differences observed in period might be explained partly by differences in expression, the other sources of variability might be attributed to different background mutations in the lines used for period rescue.

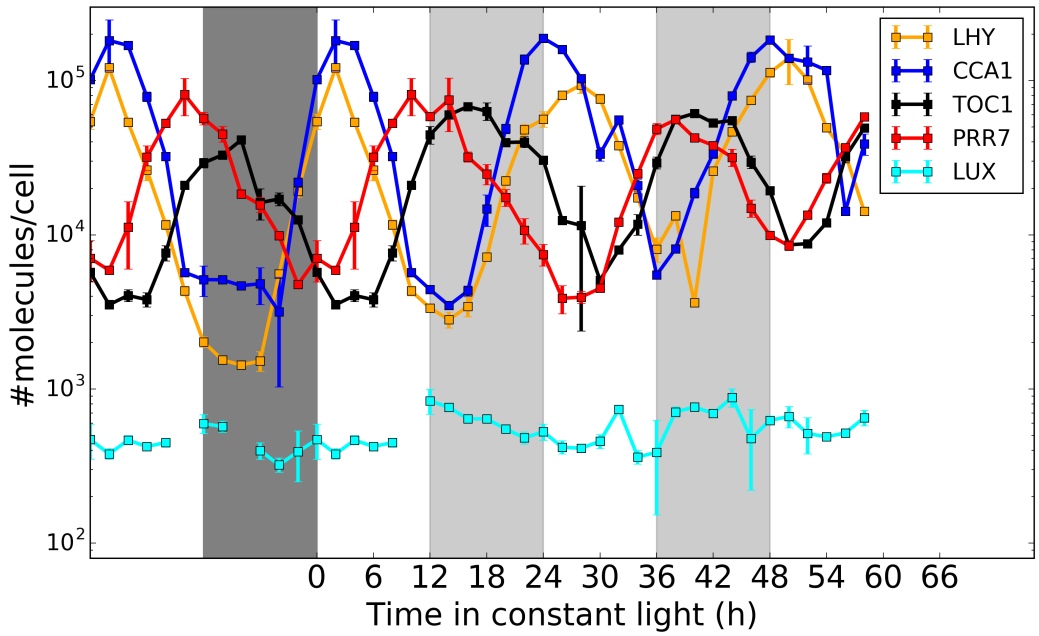


Figure 6.4 Diurnal and circadian dynamics for key clock proteins. Samples taken every two hours. Two biological replicates per time-point, three technical replicates per biological repetition. Mass-scale obtained by calibration curve using purified NanoLUC standard. Error bars show S.E.M for two biological replicates. The lines tested are indicated in table 5.1, however LUXNL21.2 was used in this experiment. The time points from ZT0-12 are the same as the first 12 hours in the plot. Dark grey bar night, light grey expected night in constant light conditions.

Quantitative features present in new proteins time series.

The following quantitative features are noteworthy. The quantification shows that the protein numbers levels can go from 300-200,000 molecules/cell for all the lines. With fold strong fold changes between peak and trough (Table 5.3). Another interesting result is the relative level between different constructs. This is not trivially obtained by selecting for high NanoLUC signal. It was an

emergent result from selecting period complementing lines using the LUC markers. The BOA-NL lines described in Chapter 4 was easily characterised though it had far low copy number. It is potentially significant, therefore, that LUX level in the LUXNL23.2 line was also low, at 100-1,000 molecules per cell. Interestingly the LUX-NL transgene at this level can complement the arrhythmic phenotype of *lux-2* in constant light conditions (Figure 5.9). The fact that these low levels are enough suggests that the TF is active even after addition of the NL tag. Also, appreciate the increase in LUX-NL amplitude in constant light (Figure 5.5).

Table 6.3 Molecules/cell at peak and trough in circadian time series for key clock proteins. Fold-change indicated All values were rounded to the floor integer.

	LHY-NL	*CCA1-NL*	*TOC1-NL*	*PRR7-NL*	*LUX-NL*
peak LD	121118.	182067	41235	81208	596
trough LD	1431	4682	3523	5886	379
fold-change	84	38	11	13	1
peak LL	139720	188712	67995	81208	882
trough LL	2816	3484	352	3875	362
fold-change	49	54	19	20	2

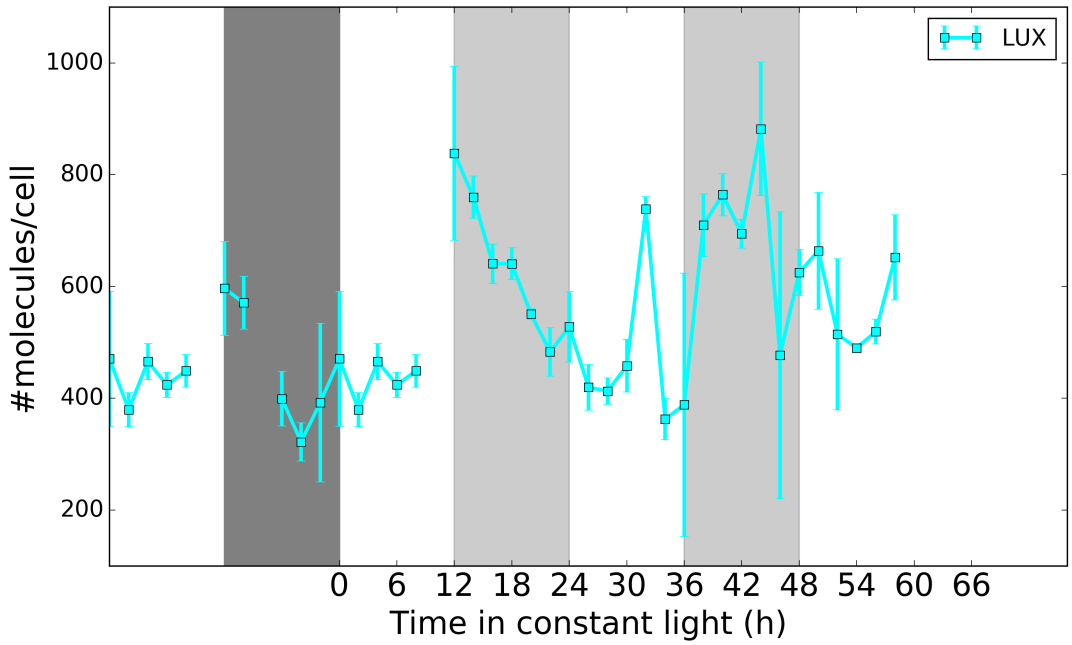


Figure 6.5 LUX-NL levels represented in a linear scale. Same data as in figure 5.4. The graph notation is the same as in figure 5.4. The amplitude of LUX increases in constant light.

Specific features for each protein can be noted thanks to this high-quality time series. For example, both LHY-NL and CCA1-NL proteins peak at ZT2 but there are indications of subtle differences between the two signals. For example, the peak time after the first day in LL differs by four hours (LHY ZT24 LL, CCA1 ZT28 LL). The trough levels are also different in the evening in the LD regime although this difference disappears in LL ZT15 LD and ZT15 LL (Figure 5.11). The two proteins rise earlier in LL ZT18 than in LD ZT18 (Figure 5.12A). The next feature is lower levels of PRR7 and TOC1 protein in darkness (ZT14 LD) compared to subjective night under constant light (ZT14-24 LL). Interestingly, CCA1 and LHY protein levels rise (ZT16 LL) before the levels of PRR7 and TOC1 decay in LL (ZT16 LL) (Figure 5.12B).

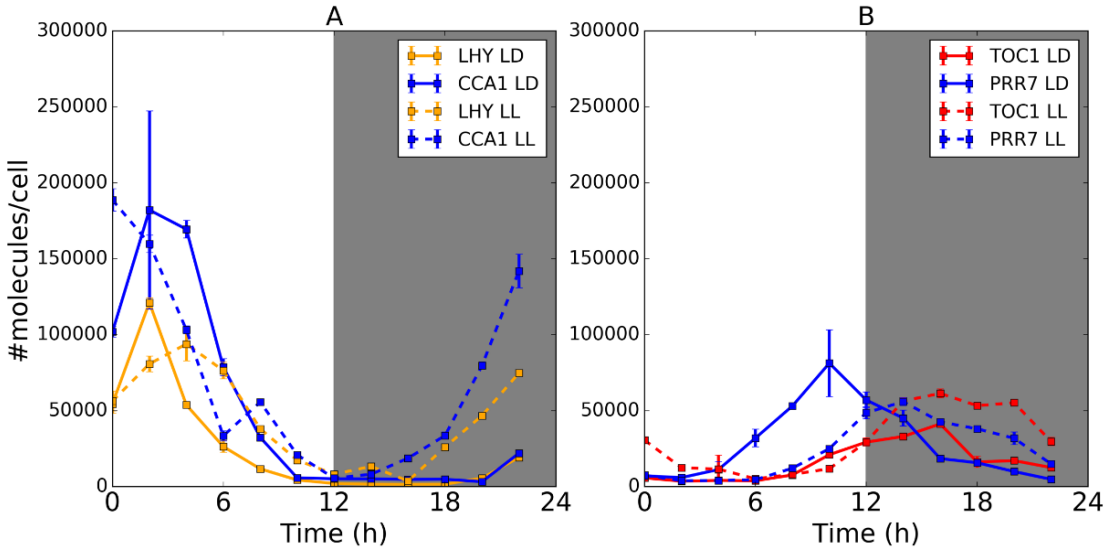


Figure 6.6 Dynamic differences between LD (ZT0-22) and LL (ZT24-48). A) LHY yellow and CCA1 blue proteins in LD solid lines and LL dotted lines. Notice early rise of LHY and CCA1 in LL. B) PRR7 blue and TOC1 red in LD solid lines and LL dotted lines. PRR7 and TOC1 peak levels change between light regimes. Data and plotting conventions as in Figure 5.11; notice the linear scale. S.E.M.

TiMet RNA qRT-PCR vs NanoLUC time-series.

In this section I compare the dynamics of a gold standard RNA time-series (Flis et al. 2015) and NanoLUC data. Appreciate that there is a difference of 2-3 orders of magnitude between qRT-PCR peak levels and peak levels for LHY-NL and CCA1-NL proteins (Figure 5.7). Also, interestingly compared to the RNA the trough levels remain above 1,000 molecules. These proteins start

rising at the same time as the RNA. However, they peak two hours later at ZT2 compared to the RNA ZT0. Interestingly, the RNA levels in constant light start rising at ZT12 while both CCA1-NL and LHY-NL protein experience a delay of 4h before they start accumulating again. This feature disappears in the second day ZT36 LL. Another feature is the peak broadening starting at ZT12 LL for both protein and RNA. In LD protein, the rise and decay spans 16h ZT while in LL it spans 22h. This is more evident in the RNA than in the protein time-series. It seems from this two independent data sets that the protein and the RNA present similar dynamics. Notice the lower amount of experimental error in the NL data compared to the qRT-PCR RNA measurements.

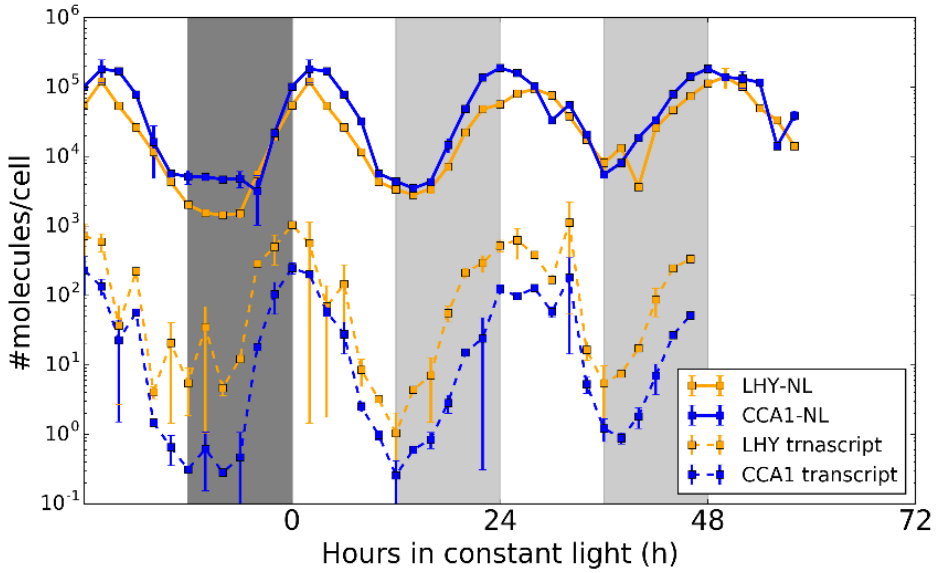


Figure 6.7 CCA1 and LHY transcript vs NanoLUC data for the same proteins. Dark grey represents darkness, light grey represents expected night in LL. Error bars are S.E.M.

The dynamics of TOC1-NL and PRR7-NL to the TiMet RNA profiles show differences in the decay of transcripts and proteins (Figure 5.8). Both transcripts peak around ZT8-12. PRR7-NL peaks at ZT10 peaks two hours after its transcript. Compared to the CCA1-NL and LHY-NL there is more noticeable mismatch between waveform of the protein and the RNA of *PRR7*. This might be a consequence of just complementing the *prr7* mutation of the *prr79* parental line. However, it is known that the PRR7 protein is subjected to strong light regulation (Farré & Kay 2007). TOC1-NL has a larger delay 6h between RNA and protein peaking time, in both LD and LL regimes. The protein rise 4h later than the RNA. Interestingly in LD conditions both PRR7-NL and TOC1-NL hit trough levels around ZT0. This is not the case in ZT24 LL. This can also be observed in *TOC1* RNA. This rises a concern regarding the regulation of CCA1 and LHY which will be discussed in further detail later on in this chapter.

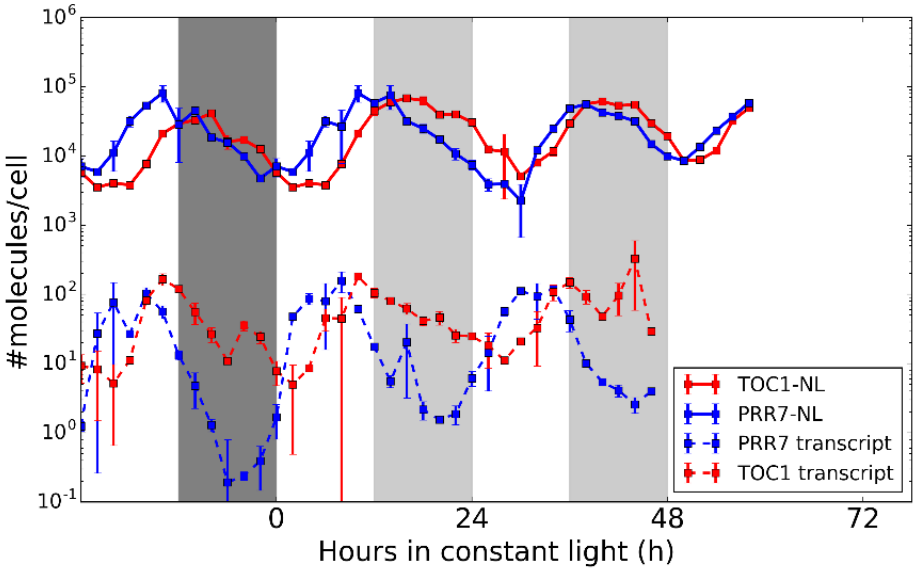


Figure 6.8 PRR7 and TOC1 transcript vs NanoLUC data, figure notation similar to Figure 6.7. Error bars: S.E.M.

Finally, *LUX* transcript present 100-fold change in LD. However, this fold change decreases slightly in LL. The base lines of the RNA increases in LL vs LD. In contrast, the quantification of LUX-NL in for line LUXNL23.2 presents lower amplitude oscillations with a very small 1.3-fold change in LD and a 2 fold-change in LL. Nonetheless the protein still present peak levels two hours later than its transcript ZT12. The low amplitude of LUX-NL can also be appreciated in the western blot data reported by (Nusinow, Helfer, Hamilton, King, Imaizumi, Schultz, Farré & Kay 2011a).

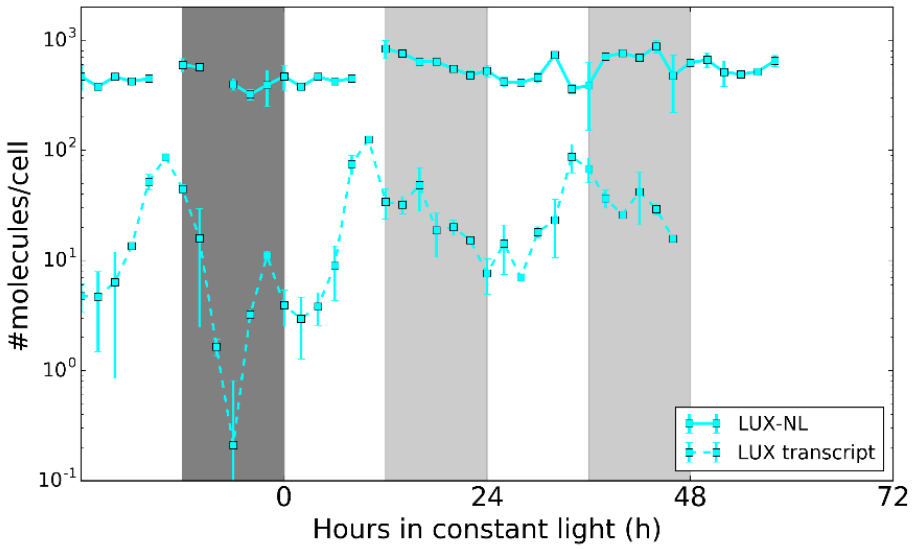


Figure 6.9 LUX trancript vs LUX-NL levels, figure notation similar to Figure 6.7. Error bars: S.E.M.

NanoLUC data vs predicted protein levels using the simple translation model.

In Chapter 4, I postulated levels of clock proteins by using the RNA Timet data and a simple mathematical model of translation. Surprisingly the predicted levels using this simple model agree in the number of molecules, apart from LUX protein. The predictions of CCA1 and LHY proteins are very interesting. In the simple model, I assume that these proteins are equivalent and therefore summed up the predicted levels. The rescue lines on the other hand have either CCA1-NL (*lhy*) or LHY-NL (*cca1*), lacking the protein in parenthesis. Both lines can complement the clock and surpassingly they complement with

similar levels of protein. Furthermore, both the prediction of the simple model and the experimental data are around the same magnitude.

Table 6.4 Predicted levels of clock proteins using a simple translation model vs NanoLUC quantifications. Prediction from simple model (pred)

	CCA1/LHY pred	*LHY-NL*	*CCA1-NL*	*PRR7 pred*	*PRR7-NL*
peak LD	106759	121118	182067	37342	81208
trough LD	68	1431	4682	1613	5886
fold-change	1569x	84x	38x	23x	13x
peak LL	70339	139720	188712	61713	81208
trough LL	4848	2816	3484	1707	3875
fold-change	14x	49x	54x	36x	20x

	TOC1 pred	*TOC1-NL*	*LUX pred*	*LUX-NL*
peak LD	23313	41235	104068	596
trough LD	493	3523	63780	379
fold-change	47x	11x	1.63x	1.5x
peak LL	29981	67995	128663	882
trough LL	452	3523	62252	362
fold-change	66x	19x	2.06x	2.4x

The simple model underestimated the levels of TOC1 and PRR7 proteins I originally suggested this in Chapter 4. The amplitude of the oscillation is higher in the predictions compared to the NL data. Nonetheless, I would say that there is a good agreement between the predictions and quantifications. However, this is not the case for the LUX-NL23.1 line selected. The predicted LUX levels are 2 orders of magnitude off. However, the fold change is remarkably similar. Suggesting that the degradation rate of the LUX protein fused to either GFP (Nusinow, Helfer, Hamilton, King, Imaizumi, Schultz, Farré & Kay 2011b) or NanoLUC is low. The RNA presents a strong fold change which is not the case for the protein. I performed exploratory plate reader experiment with LUX-NL9.2 line which has a closer period to the WT CAB2p:LUC marker (Table 6.2). The gold standard data for CCA1-NL and TOC1-NL allows introducing a rough absolute scale for the plate reader experiment (Figure 6.10). This experiment shows that levels of the secondary line are higher, however they are below 10,000 suggesting that the clock can still work with very low level of LUX and possibly the rest of LUX protein might be involved in output pathways (LUX-NL 21.2 have longer hypocotyl compared to the WT and closer to the mutant phenotype while LUX-NL 9.2 closer hypocotyl length to the WT). The amplitude of the oscillation is low in both lines compared to the other measured proteins though (TOC1-NL and CCA1-NL).

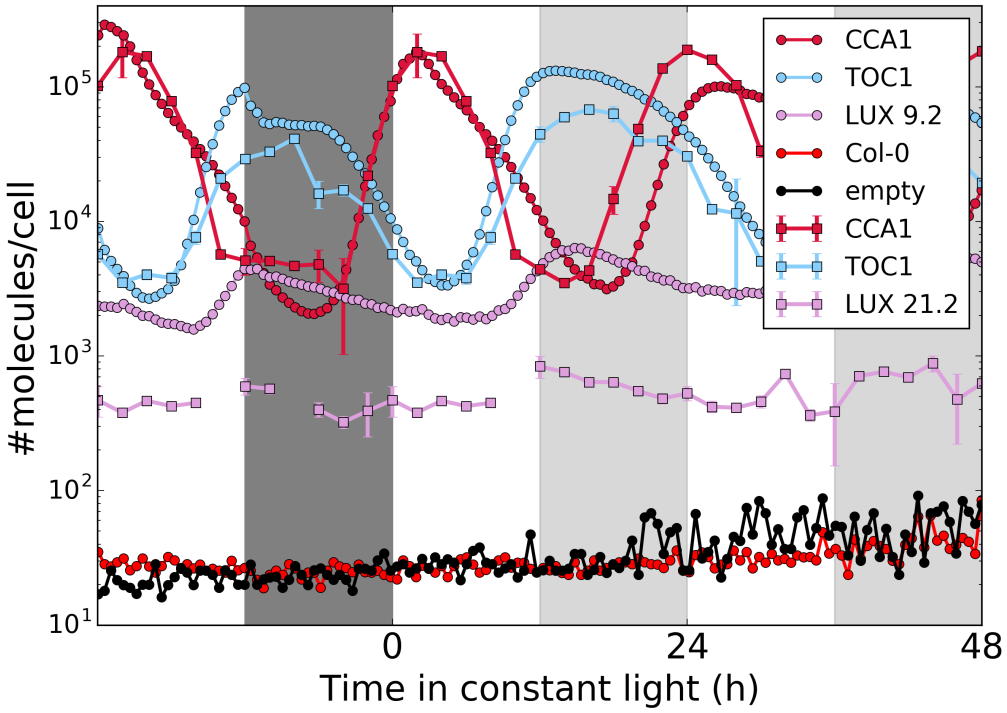


Figure 6.10 Plate reader experiment for characterizing the levels of a secondary LUX-NL line. CCA1-NL and TOC1-NL from table 6.2 were used. LUX-NL 21.2. Background signal in Col-0 plants and empty well is similar. The gold standard in squares and circles plate reader. The plants were grown for two weeks in ROBUST agar then transferred to 96-well black well plate with ROBUST media with 1:50 Furimazine:0.01% Triton X-100. Light conditions are the same as described in the Chapter 5. Error bars: S.E.M.

U2017.3 vs protein data in absolute units for LHY and CCA1.

In chapter 4, I rescaled the protein component of U2017.1 resulting in U2017.3 by using the protein predictions from the simple model. CCA1/LHY has smaller amplitude in U2017.3 relative to the data in both LD and LL. The model presents a lower fold change in LL compared to the NanoLUC data. The amplitude of CCA1-NL and LHY-NL is maintained in constant light, which is different to U2017.3 predictions (Figure 6.11).

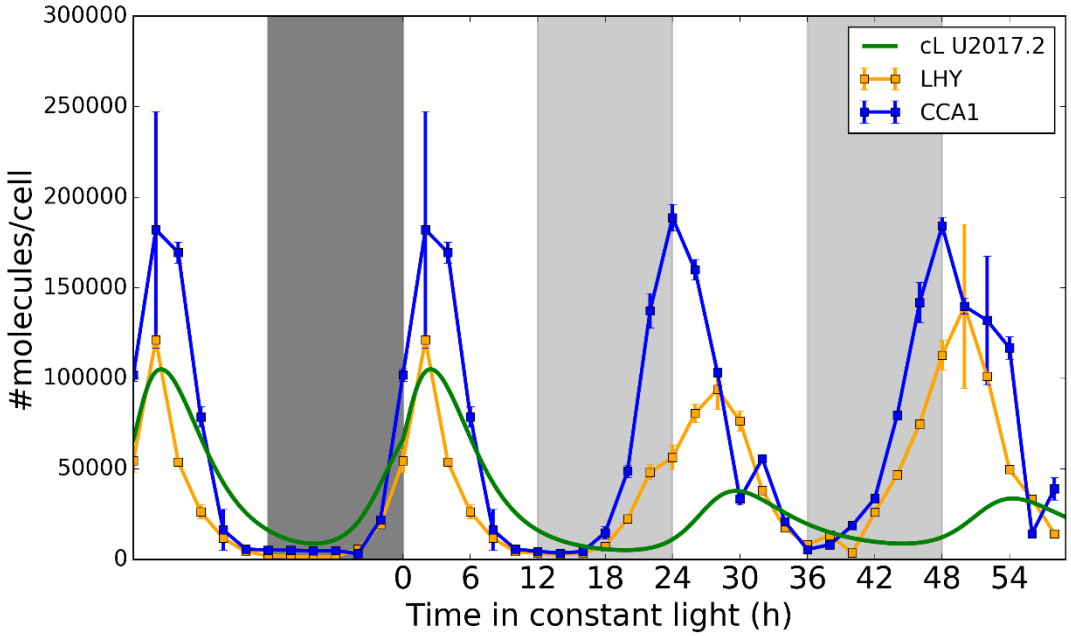


Figure 6.11 Comparing U2017.3 to absolute quantification of LHY-NL and CCA1-NL. Quantitative comparison is summarised in Table 6.5. Error bars: S.E.M.

Table 6.5 Predicted levels of clock proteins using U2017.2 which was absolute number of clock proteins vs NanoLUC data.

	CCA1/LHY pred	*LHY-NL*	*CCA1-NL*	*PRR7 pred*	*PRR7-NL*
peak LD	104974	121118	182067	46713	81208
trough LD	8748	1431	4682	1419	5886
fold-change	11x	84x	38x	32x	13x
peak LL	33625	139720	188712	16447	81208
trough LL	8742	2816	3484	3213	3875
fold-change	3x	49x	54x	5x	20x

	TOC1 pred	*TOC1-NL*	*LUX pred*	*LUX-NL*
peak LD	6066	41235	147927	596
trough LD	1170	3523	17236	379
fold-change	5	11	8	1
peak LL	14951	67995	164436	882
trough LL	1661	3523	23471	362
fold-change	9x	19x	7x	2x

U2017.3 vs protein data in absolute units for PRR7 and TOC1.

I selected PRR7 and TOC1 proteins for quantification in order to test the relative levels of these proteins predicted by U2017.3. In the model, the transcriptional regulation of *CCA1/LHY* enforces the same scaling factor for all the PRRs, therefore fixing the relative levels between them, described in Chapter 4. In the model, the peak levels in LD and LL are more different compared to the trough levels. In the NanoLUC data the peaks are more similar (Table 6.6). Notice in the time series that in the model PRR7 protein decays much faster in darkness than the PRR7-NL. Also, the levels or this protein are lower in darkness than in constant light expected night ZT12-24 LL. TOC1-NL present also higher levels in the expected night in LL compared to darkness. Interestingly high levels of TOC1-NL extend beyond ZT24 LL (Figure 6.12).

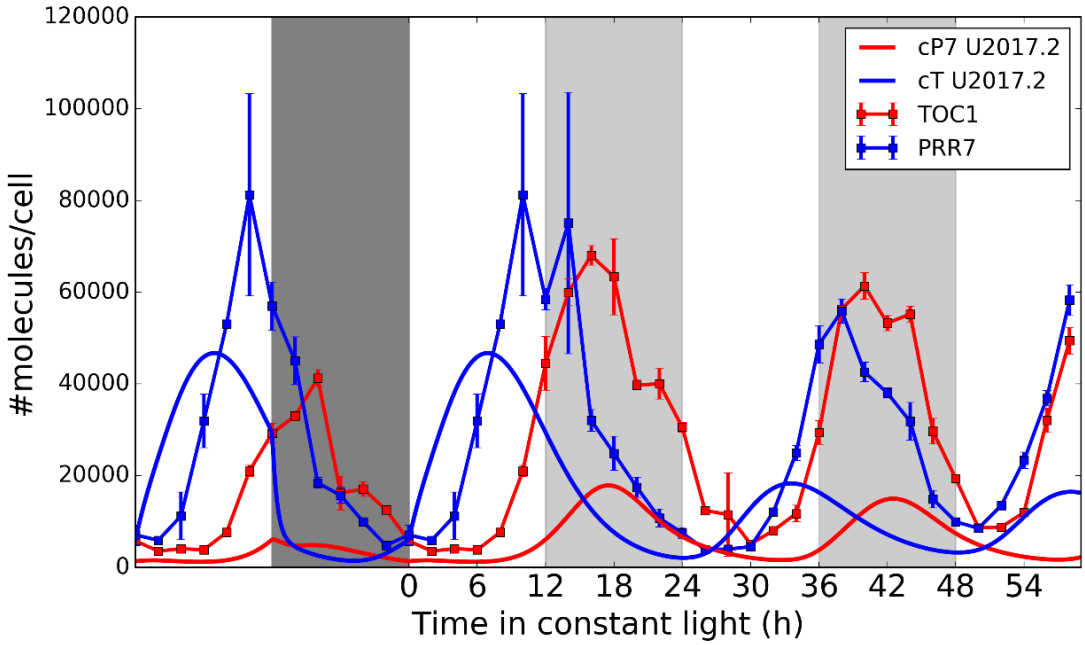


Figure 5.12 U2017.3 simulations vs NL protein data in absolute units for PRR7 and TOC1. The model u2017.2 entrained for 10xLD cycles and the released into LL. Table 6.6 surmises the relative levels of PRR7/TOC1 n the model. Error bars S.E.M.

Table 6.6 Relative levels of PRR7/TOC1 in U2017.3 and NanoLUC data.

	PRR7pred/TOC1 pred	*PRR7-NL/TOC1-NL*
peak LD	6.15	1.96
trough LD	1.37	1.67
peak LL	4.12	1.19
trough LL	1.02	1.10

Phase plots of TOC1 protein vs *LHY and CCA1 transcript reveal extra complexity of PRR regulation.*

TOC1 has been proposed, along with other PRRs, as a repressor of *CCA1* and *LHY* transcription (Alabadí et al. 2001)(Gendron et al. 2012) (Huang et al.

2012). Thus, I would expect that at high levels of TOC1 protein, low levels of *LHY* and *CCA1* transcript synthesis would be present. To better visualize this, I created a phase plot between the dynamics of these elements using the NL protein and TiMet RNA data (Figure 6.13). During the rise region of TOC1-NL protein level increase is correlated with a decrease in the levels of *LHY* and *CCA1* transcripts, both in LD (ZT8-16) and LL (ZT8-16) conditions. This is not the case when the levels of TOC1 are decreasing. Interestingly the levels of *LHY* and *CCA1* transcripts rise before TOC1-NL protein levels drop to the equivalent levels that appear to suppress the transcripts in their rising phase. There is some resemblance of this behaviour in the U2017.2 model but the lack of activity of TOC1 protein (cT) on *LHY/CCA1* (*cLm*) is more evident in the data. It seems that the anti-correlation between TOC1 protein levels and *LHY* and *CCA1* transcript levels is lost in constant light conditions in the decay phase of TOC. I will discuss a possible mechanism for solving this paradox because *LHY* and *CCA1* promoters are canonical targets of TOC1 transcriptional activity.

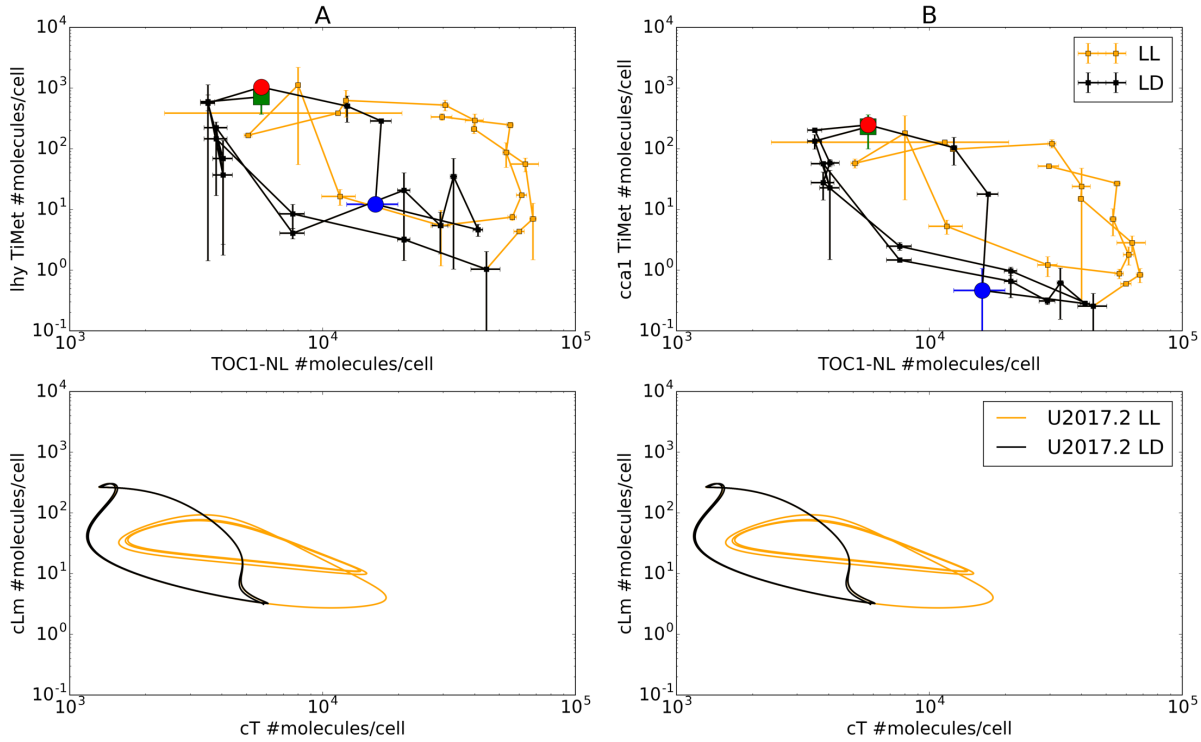


Figure 6.13 Phase plots of CCA1 and LHY transcripts from TiMet vs TOC1-NL. A) Upper panel shows TOC1-NL vs LHY transcript. In black the LD regime. The staring time point is shown in green the clock the clock progression is counter clock wise. The end of the first LD day is marked with a red circle. Notice the rise of transcript levels at ZT10 blue circle. The LL regime is depicted in orange. The model U2017.2 is plotted in the lower panel. The model can explain some of the response, however the data has a wider limit cycle in constant light LL. Errors bars: S.E.M.

Absolute quantification of LUX reveals unrealistic function of the EC.

It is remarkable that only low levels of LUX protein were required to rescue the arrhythmic phenotype of *lux-2* earlier in this chapter. This has important consequences for the model. In U2017.2, cLUX simulated protein levels are on the order of $1x10^5$ molecules per cell at peak times. Originally U2017.2 proposed a very slow formation of the evening complex, which results in the

order of 10,000 cEC molecules at peak times (Figure 5.14). If the model is rescaled to match the LUX-NL data instead of the synthetic protein data that would result in even lower levels of cEC. In turn, this will require affinity way beyond the nM range. The implications of this result relate multiple aspects of the thesis and are considered in the final discussion.

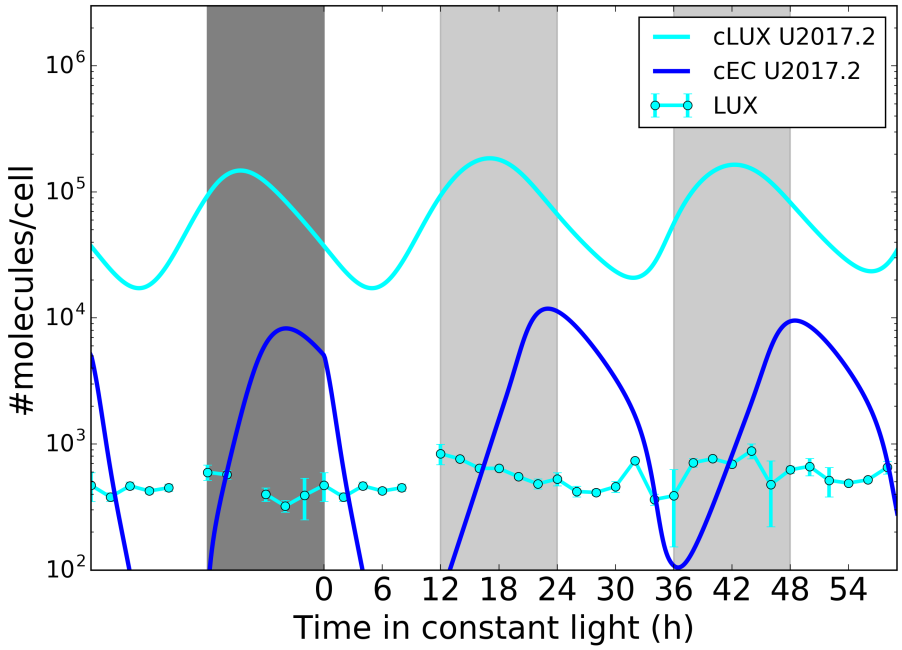


Figure 6.14 Predicted levels of LUX-NL suggest incorrect peaking phase of EC formation. Error bars: S.E.M.

Discussion

The implication of the data of the data in the model were considered in the above results. While this data links all the chapters as it builds up on each other to arrive the results presented here I will summarise and discuss the links

In the General Discussion Chapter 7. Therefore, this discussion focuses mainly in technical aspects regarding the absolute quantification.

In this chapter, I presented the extensive work required for testing the prediction of protein levels performed with the simple model from Chapter 4. These new lines will facilitate a highly quantitative exploration of the clock at the protein level for the first time in *Arabidopsis*. However, further characterisation of this lines is required regarding to test the equivalence of the transgene to the original proteins. For example, it is possible that the activity of the native protein is compromised both decrease in activity or increase in activity of the protein would result in a wrong estimation of the native protein levels. Therefore, other tests for this modified version might be needed for example EMSA assays showing that the affinity for known binding sites is not compromised. I based my designs in previous knowledge of tagged lines for the same proteins. Therefore, problems associated with these lines might be also reflected in the other lines tagged lines. However possibly the first indication of lack of proper rescuing could be differences between the levels transgene and the WT levels. If the RNA levels are very similar between the transgene and the WT levels it could indicate that the protein retained its activity. On the other hand, different levels could indicate hyperactivity or reduced activity in the tagged protein. Another important test is the profile of the RNA because I modified the 3'-UTR by introducing the NanoLUC and a

standard NOS-terminator. Therefore, further characterisation is required to test if the circadian dynamics are retained at the RNA level for the transgene.

More work is required for the double mutants *lhy/cca1* and *prr7,9*. In these two cases, the secondary mutation needs to be eliminated by crossing each collection to the single mutant. Nonetheless, this first step of rescuing the double mutants allowed me to clearly see the strong recovery from the phenotype observed in the double mutants. For example U2017.2 underestimates the levels of LHY and CCA1 protein. However, it has been shown that the affinity of CCA1:LHY dimer is higher than proteins working independently (O'Neill et al. 2011). Therefore, it is possible that a higher amount of protein is required when the period is rescued by only one of them. In the case of PRR7-NL all the lines presented periods close to 26h or longer. This suggest that the PRR9 proteins might be required for full period complementation. This might have a final effect resulting in better quantifications.

Given the limited time-span for this part of the project, I focused in the period phenotype but further characterisation is required for other phenotypic markers. Two phenotypes that are generally tested are hypocotyl length and flowering time. In particular characterising, them this will be important for improving the quantifications. In my experience working with the plants they present variability in the latter traits. In particular the extremely late flowering

time of *prr7,9* was rescued (Figure 6.15). However, very careful analysis for these traits is required in the transgenic collection which con be done in follow up work.

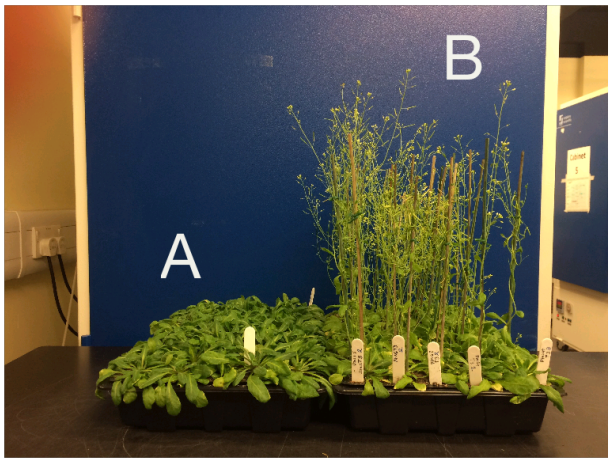


Figure 6.15 PRR7-NL rescues the late flowering time of prr7,9. A) prr7,9 parental lines. B) T3 transgenic lines prr7,9 PRR7-NL grown in 16L:8D photoperiod 100 µmol/cm^{2}s^2 white warm light ~6 weeks old plants.

I showed that in addition of performing manual sampling experiment. It is also possible to perform an automatic and in plant tracking of protein levels at a higher temporal resolution than ever before using plate readers. This duality allows performing circadian exploratory experiments before manual sampling. The protocol I developed for performing absolute quantification is robust and provides protein data with comparable quality to qRT-PCR experiments. Also, the dynamic range of the bioluminescence assay allows an easier way to avoid saturation issues. However, better characterisation of the experimental error

should be performed. This is important if the data if fed into clock model fitting as it was discussed in Chapter 3 for the RNA data.

References

Alabadí, D. et al., 2001. Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science (New York, N.Y.)*, 293(5531), pp.880–883.

Corellou, F. et al., 2009. Clocks in the green lineage: comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. *The Plant cell*, 21(11), pp.3436–3449.

Farré, E. & Kay, S., 2007. PRR7 protein levels are regulated by light and the circadian clock in Arabidopsis. *The Plant journal : for cell and molecular biology*, 52(3), pp.548–560.

Farré, E.M. et al., 2005. Overlapping and Distinct Roles of PRR7 and PRR9 in the Arabidopsis Circadian Clock. *Current Biology*, 15(1), pp.47–54.

Flis, A. et al., 2015. Defining the robust behaviour of the plant clock gene circuit with absolute RNA timeseries and open infrastructure. *Open biology*, 5(10), p.150042.

Gendron, J. et al., 2012. Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8), pp.3167–3172.

Hansen, L.L. & van Ooijen, G., 2016. Rapid Analysis of Circadian Phenotypes in Arabidopsis Protoplasts Transfected with a Luminescent Clock Reporter. *JoVE (Journal of Visualized Experiments)*, (115), pp.e54586–e54586.

Hazen, S.P. et al., 2005. LUX ARRHYTHMO encodes a Myb domain protein essential for circadian rhythms. *Proceedings of the National Academy of Sciences*, 102(29), pp.10387–10392.

Huang, W. et al., 2012. Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator. *Science (New York, N.Y.)*, 336(6077),

pp.75–79.

James, A.B. et al., 2012. Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *The Plant cell*, 24(3), pp.961–981.

Kamioka, M. et al., 2016. Direct Repression of Evening Genes by CIRCADIAN CLOCK-ASSOCIATED1 in the Arabidopsis Circadian Clock. *The Plant cell*, 28(3), pp.696–711.

Lee, C. et al., 2001. Posttranslational mechanisms regulate the mammalian circadian clock. *Cell*, 107(7), pp.855–867.

Mombaerts, L., Mauroy, A. & Gonçalves, J., 2016. Optimising time-series experimental design for modelling of circadian rhythms: the value of transient data. *IFAC-PapersOnLine*, 49(26), pp.109–113.

Narumi, R. et al., 2016. Mass spectrometry-based absolute quantification reveals rhythmic variation of mouse circadian clock proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24), pp.E3461–7.

Nusinow, D., Helfer, A., Hamilton, E., King, J., Imaizumi, T., Schultz, T., Farré, E. & Kay, S., 2011a. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

Nusinow, D., Helfer, A., Hamilton, E., King, J., Imaizumi, T., Schultz, T., Farré, E. & Kay, S., 2011b. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356), pp.398–402.

O'Neill, J. et al., 2011. Circadian clock parameter measurement: characterization of clock transcription factors using surface plasmon resonance. *Journal of biological rhythms*, 26(2), pp.91–98.

Seo, P.J. et al., 2012. A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in Arabidopsis. *The Plant cell*, 24(6), pp.2427–2442.

Shen, H., Moon, J. & Huq, E., 2005. PIF1 is regulated by light-mediated degradation through the ubiquitin-26S proteasome pathway to optimize photomorphogenesis of seedlings in Arabidopsis. *The Plant journal : for cell and molecular biology*, 44(6), pp.1023–1035.

Strayer, C. et al., 2000. Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. *Science (New York, N.Y.)*, 289(5480), pp.768–771.

van Kregten, M. et al., 2016. T-DNA integration in plants results from polymerase-θ-

mediated DNA repair. *Nature Plants*, 2(11), p.16164.

Yoo, S.-H. et al., 2004. PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. *Proceedings of the National Academy of Sciences*, 101(15), pp.5339–5346.

Chapter 7
General Discussion

Considering the dimensionality of the system is an important consideration when starting to model a system. In particular for molecular networks determining the system size will help to determine if a deterministic approach is the best way for modelling the system. If low number of molecules will result in stochastic effects that might have important consequences for the dynamics. Even if the preferred approach is a deterministic approach it might be possible that the functional forms in the model might change depending in the information available. For example, in oscillators nonlinearity contribute to robustness in oscillations without entrainment. Historically for the circadian oscillator, hill functions and coefficients have been exploited for bringing ultra-sensitivity into the dynamics. However, the use of absolute units allows consideration of number of sites in the genome which might act as decoy elements titrating the transcription factors (Buchler & Cross 2009). If the concentrations and affinities are in particular ranges, then ultrasensitive behaviour with very high hill numbers can be achieved. In the context of absolute units, the effects of crowding can start to be evaluated and how it impacts the occupancy of transcription factors to the genome (Kemme et al. 2015). Nonetheless even though it is desirable to provide absolute units

generating this type of data is very demanding and requires very highly skilled work which might complicate the access for theoretical researchers. Therefore, improvement in methodologies that provide this type of dynamic data in an automatic way needs to be promoted.

The value of absolute units for plant clock models

In this thesis, my aim was to recast the Arabidopsis clock model from arbitrary to absolute units, in order to test for assumptions or kinetic parameters in the existing models that were not biochemically realistic. Earlier chapters described several examples of this process and further remedies are discussed below. possibly the most remarkable result is how the use of absolute units is valuable for integrating data in a way that was not possible using models with arbitrary units. It is clear that the TiMet RNA data provides a Gold standard for testing current and new clock models. However, the use of absolute units impacts beyond the improvement of the oscillator models. It allows unification of quite different data types. Furthermore, I showed that mathematical modelling is central for this unification process.

The TiMet data was generated for testing P2011. It has a good sampling rate and informative clock dynamics (transition from LD into LL, and into DD). It eliminates the use of internal standards for controlling RNA extraction, which can result in artefactual circadian waveforms, in turn compromising model

assessment and parameter fitting. Therefore, eliminating this normalisation issue is a major experimental step forward in testing clock models. This data alone is not enough for assessing un-realistic biochemical rates in the model. Additional information is required that draws the lines between biologically realistic or parameters out of range. The original approach I took was to borrow data from other eukaryotic systems to obtain maximum transcription rates for each the clock genes. However, I found an alternative approach consisting in linking two published RNA data sets through (Piques et al. 2009; Sidaway-Lee et al. 2014) simple mathematical modelling (Chapter 3). The result was a more natural, empirical probability distribution of transcription rates in units of [transcripts]/[cell h]. The distribution could be a valuable tool not only for circadian models but other transcriptional regulation models or in parameter inference methodologies. For example, in Bayesian approaches it can be used as a prior distribution representing what we know about transcription when inferring model parameters.

In addition to improving the RNA dynamics of the model, the TiMet RNA data was fundamental for producing predictions about protein levels of clock genes. RNA data has been used extensively as a readout of the clock's internal state. Even tough transcription factors are currently the main players in the network. Little protein data exists for them. Predicting the amount of protein from RNA data allowed me to explore the implications of absolute protein units for two years, while producing the stable transgenic lines that

would allow me to quantify them experimentally. Having the predicted numbers in hand uncovered the need of integrating genomic data (Chip-seq), TF sequence preference (PBMs) and cell compartment volumes to link binding data (SPR) to the mechanistic model U2017.1. This linking allowed me to assess if the affinities for single sites are realistic in the model. However, similar to the transcription rates case, information about the distribution is important for assessing these values. I collected a set of parameters from the literature resulting in a first distribution for TF:DNA interaction. Similar to the distribution of transcription rates in absolute units, this distribution can the TF constrain binding affinities in model optimisation, which will be valuable for Bayesian inference. However only around 1% of the *Arabidopsis* TFs affinity to DNA have been studied directly. Possibly, many of them are still buried in the literature. Therefore, further bioinformatics work should be performed for cataloguing them. Nonetheless, the shape of the distribution can be approximated by a log normal model, this has been observed for other biological parameters like enzymatic activities from the Brenda database (Bar-Even et al. 2011)

From the theoretical exercise of hypothetic proteins levels, I derived a set of genes for generating stable transgenic lines fuses to NanoLUC. Therefore, focusing the efforts into those genes that were particularly informative and amenable to experimental investigation. Perhaps surprisingly, my experimental results show that the synthetic protein levels, predicted from

TiMet RNA data, were on the correct order of magnitude, strengthening the results of both approaches (theoretical and experimental). Furthermore, the protein quantification provided further evidence that TiMet project provided good estimation of RNA levels, reinforcing these results and my protein measurements.

This thesis overall shows how absolute units provide a unifying framework for deepening our understating of the clock mechanisms. It also suggests how the clock model can be extended into higher organismic scales. This could not have been achieved by just exploring new architecture for gaining better fits to RNA time series. The use of absolute units in the model will be fundamental as we move from the clock network to transcriptome wide dynamic models.

In the following sections I will discuss possible ways of model improvement based on the results presented in the results chapters. I will draw a roadmap for achieving transcriptomic wide models, or for linking clock models to natural variation measured by the Arabidopsis 1001 genomes project.

Uncovering a possible role for PRR phosphorylation.

One result from rescaling the protein components was uncovering that the translation rates for the PRRs are possibly unrealistic. The experimental

measurements suggested that the translation rates in U2017.2 are even higher than anticipated by the synthetic data for PRR7 protein. The conundrum is that the decay rates of the PRR proteins need to be high, so that their decay allows *LHY/CCA1* transcript to present the correct phase of expression. Darkness allows rapid destabilisation of the PRR proteins in the model, however this effect is lost in constant light conditions. Interestingly my experimental measurements for PRR7-NL and TOC1-NL suggest that *LHY/CCA1* RNA levels can be decoupled from the PRR protein levels. Even at high levels of PRRs, the *LHY* and *CCA1* transcripts rise before subjective dawn. A possible explanation for this is that two pools of PRR proteins exist, a clock-active pool and a clock-inactive one, while the reporter fusion measures both pools. It is possible that the clock-inactive pool covers other requirements in the plant physiology that are clock independent. Partitioning of the PRRs could be achieved through some sort of post-translational modification. It has been shown that the PRRs are phosphorylated in both LD and LL conditions in a circadian manner (Fujiwara et al. 2008). The PRRs recruit the co-repressor TOPLESS (TPL) to the promoters of *CCA1* and *LHY* at ZT6 and not at ZT18 (L. Wang et al. 2013). TPL is present across the day suggesting that its repressive activity could be recruited at any time of day. This could mean that in constant light the PRRs should produce stronger repression on the promoter of *CCA1* and *LHY*. This is opposite to what my NanoLUC data shows. *LHY* and *CCA1* transcripts rise almost irrespective of

the levels of TOC1 in LL. However, if phosphorylation inhibits the interaction PRRs:TPL this could explain the timing of *LHY* and *CCA1* transcript.

These observations suggest the introduction of hypothetical kinase in the model that presents circadian activity acting towards the end of the subjective day on the PRRs. the WNKs (Kinases With No Lysine (K)) present evening to mid-night phase of expression with robust transcript circadian oscillations in LL (Nakamichi et al. 2002). It has been shown that WNK1 kinase can phosphorylate PRR3, but no *in vitro* activity was reported for the other PRRs, nor for other WNK kinases. This does not rule out the possibility for low enzymatic activity or the need for other elements for proper kinase activity. WNK1 also interacts with PRR5 in yeast two hybrid experiments (Murakami-Kojima et al. 2002). Independent work showed that PRR5 and TOC1 interact promoting phosphorylation and promoting nuclear import of TOC1, this interaction is mediated through the N-terminal PRR domain (L. Wang et al. 2010). Therefore, I want to suggest that the some of the PRRs act in synergistically *in-planta* for promoting phosphorylation of PRR clock activity therefore allowing decupling of the PRRs pool in two sets.

An interesting question is why PRR activity might need to be inhibited while maintaining protein levels. In principle, the proteins could be made unstable enough to act only at the correct time for repressing transcription of *LHY* and

CCA1. However, If the PRRs work in other pathways apart from the oscillator, then certain levels might be required to meet this non-clock demand. In particular, it has been shown recently that the PRRs stabilise CONSTANS (CO) protein (Hayama et al. 2017). Therefore, phosphorylation of the PRRs might help to decouple the PRRs' role in the oscillator from their role in the photoperiodic pathway of flowering. CONSTANS(CO) is an important transcriptional activator of *FLOWERING LOCUS T (FT)*. FT protein is the florigen responsible of promoting transition of the shoot apical meristem (SAM) from vegetative growth into reproductive growth. In *Arabidopsis*, genetic perturbations that affect FT levels result in flowering time phenotypes. Interestingly the *wnk1-1* presents a very late flowering phenotype (Y. Wang et al. 2008). However, plenty of further experimental work is required to test this mechanistic model uncovered by using absolute proteins quantifications.

Strengthening the role of BROTHER OF LUX ARRHYTHMO (BOA) in the clock mechanism.

One focus in the development of the F2014 model was modifying how transcriptional activation occurred in the clock, including the way in which the PRRs are regulated. F2014 also incorporated BOA as a new element that

could participate in the Evening complex (Fogelmark & Troein 2014). Fogelmark and Troein reported 8 parameter sets obtained by fitting the model to an extensive collection of time series collected from literature images. However, exploring the effect of BOA over-expression on the period of the clock in F2014 results in very different period responses among parameter sets as shown in Chapter 4. This provides evidence that F2014 is over-fitting the data and it precludes the study of BOA in this model. One concern is that there so many new elements in F2014 that it is very difficult to understand where this issue might arise. Therefore, starting from a better-understood model like U2017.2 for introducing BOA is easier and exploring its effects in this context might be a better approach. In particular adding BOA to U2017.2 can replicate qualitatively the trends observed in BOA genetic perturbations. In particular, the data I generated using BOA-NL fusions reinforced that increase of BOA expression results in period lengthening in Col-0 CCA1p:LUC lines, verifying in an independent way the reproducibility of the Dai et al results. An increase in CCA1p:LUC signal suggested that BOA promotes *CCA1* transcription, as previously reported.

Also, the presence of fusion signal at ZT19 in LD and in LL provides evidence that BOA accumulates at the required time of the day to repress the EC's clock targets. BOA has been reported to interact with ELF3 at ZT12, providing evidence for at least two versions of the EC (EC:LUX and EC:BOA).

An important draw back in analysing BOA's role is the lack of a knockout (KO) mutant. The *boa-1* T-DNA line partially affects the behaviour of the clock with a subtle dampening of *CCA1* transcript (Dai et al. 2011). On the other hand, the RNAi lines of *BOA* do decrease the RNA levels of *BOA* but no period phenotype was observed compared to WT. The LUX/BOA amiRNA performing a simultaneous knock down might mask the effect of BOA if both proteins participate in a complex. Interestingly, all the lines that I generated for LUX-NL present rhythmic oscillations for CAB2p:LUC providing evidence of rescue, however further replication of this experiments is required for properly assessing the oscillatory robustness of these lines. However, when I was selecting the T3 generation for absolute quantification, I observed important variability in flowering times with many lines presenting early flowering compared to the WT control (data not shown), similar to the *lux-2* mutant. This suggests that many of the LUX-NL are producing low levels of LUX that are not enough for complementing the output control but enough to produce oscillation in the CAB2p:LUC output. Suggesting that arrhythmia will only be observed in the *BOA* KO, not in knock-down lines that retain some *BOA* function. Thanks to the development of CRISPR this can be explored by targeting *BOA* with specific gRNAs. Evidence for targeting members of the evening complex with such an approach has been provided in duckweed with successful results for *ELF3* (Okada et al. 2017).

A different path of model improvement.

F2014 proposed a new network for the clock and obtained new parameters by fitting to an extensive collection of time series in arbitrary units. An example where the model could be exploited is by attempting to infer the extra amount of BOA being produced by the super transformation of BOA-NL I performed in Chapter 5 during the test of the NanoLUC system. In that particular scenario, a dose response could be observed on the period. Therefore, I first thought of using F2014 for inferring how much more protein these plants could be producing. F2014 reported 8 parameter sets from their model fitting. Introducing the suggested perturbation resulted in wide variation in the period change. In this particular situation, the model was of little use.

Instead of introducing more model complexity in this project, I opted for increasing the quality of the protein data bringing absolute units at that level for the first time. This provides a different path for suggesting modifications that might improve the mathematical model. This are introduction of BOA as a secondary activity in the EC and phosphorylation of PRRs for total levels for decoupling clock and output pathways and the change in the backward inhibitory wave in the regulation of these proteins.

Exploring the use of new lines for performing quantitative work at the protein level.

The results of absolute quantification showed that the NL tag can be used for performing absolute quantifications. The random integration of the synthetic constructs in the genome provides a collection of lines with different expression levels that need to be characterised in detail in further work. These new lines offer a set of tools for studying in a highly quantitative way the response of several Arabidopsis traits to varying levels of expression of clock proteins. An example would be to explore what is the response of the clock period to increasing levels of protein. This dose response effect has been studied for ZTL and for CCA1, but this has been performed with 35S lines (Somers et al. 2004; Z. Wang & Tobin 1998). My collection of homozygous transgenics should present overexpressing levels at a correct time of day. The possibility of performing absolute quantification provides a way to link this experimental perturbation analysis to clock models by simulating equivalent perturbations and comparing the trait response of data and models e.g. period change with concentration increase. Therefore, therefore characterising the sensitivity of the clock transcriptional perturbations.

Additionally, these lines can be used for optimal experimental design and parameter inference on the fly, using methodologies like the Kalman-filter. In particular the set chosen LHY/CCA1, PRR7, TOC1 and LUX allows the use of the C2016 model as a first approximation to the problem, since all the

proteins are observable in these transgenic lines. Furthermore, the period of the clock can be quantified independently of the protein marker, as each line provides firefly LUC markers (LHY/CCA1 CCA1p:LUC, PRR7 CCR2p:LUC, TOC1 CCR2p:LUC, LUX CAB2p:LUC), allowing the possibility of tracking both protein level and transcriptional dynamics.

Linking models to the genome and beyond

The use of absolute units might allow linking the model to genome level dynamics. By combining protein binding microarray data, degradation rates and transcription rates in absolute units, we can start building models with little unconstrained parameter inference or none at all. This will be an important aid for experimental work as it might help in understanding the contribution of the clock to the observed dynamics of the Arabidopsis transcriptome. A broad range of experimental transcriptomic data has indeed been published.

In addition to the regulation of single genome. I can foresee the use of these mathematical models for understanding the role of variability in promoter regions of canonical elements in the clock in the context of the 1001 genomes project. In this big data project is it is possible that mechanistic modelling will provide the integration required to for exploring the consequences of the variability observed. For example, consider the

promoter region of *GI* for all the 1001 Arabidopsis genomes. By using U2017.2 (Absolute units for all the Transcription factors), volumetric data, ChIP-seq data and a simple model of cL and cEC regulation for these 1001 *GI* promoter sequences, I observe something remarkable (Figure 7.1). First, assigning the affinities from their varying sequences resulted in a rhythmic prediction in each case, which was not guaranteed. In 12L:12D photoperiods, all the simulated *GI* RNA profiles peak at ZT12 (as expected from light regulation of *GI*) but at different peak levels. Surprisingly, when the photoperiod is modified, I uncovered that the ecotypes present a modulation of *GI* peak times. Therefore, differential affinity of CCA1 or LHY to the promoter of *GI* could result in different peaking times of *GI*, which is a key regulator of flowering time. Interestingly it has been observed experimentally that *GI* has different peaking times among accessions, depending on their geographical origin (de Montaigu & Coupland 2017). This might be explained in part through CCA1/LHY affinity for the promoter of *GI*. It is possible that mathematical models will prove to be useful for interpreting the impact of certain SNPs in the behaviour of the system.

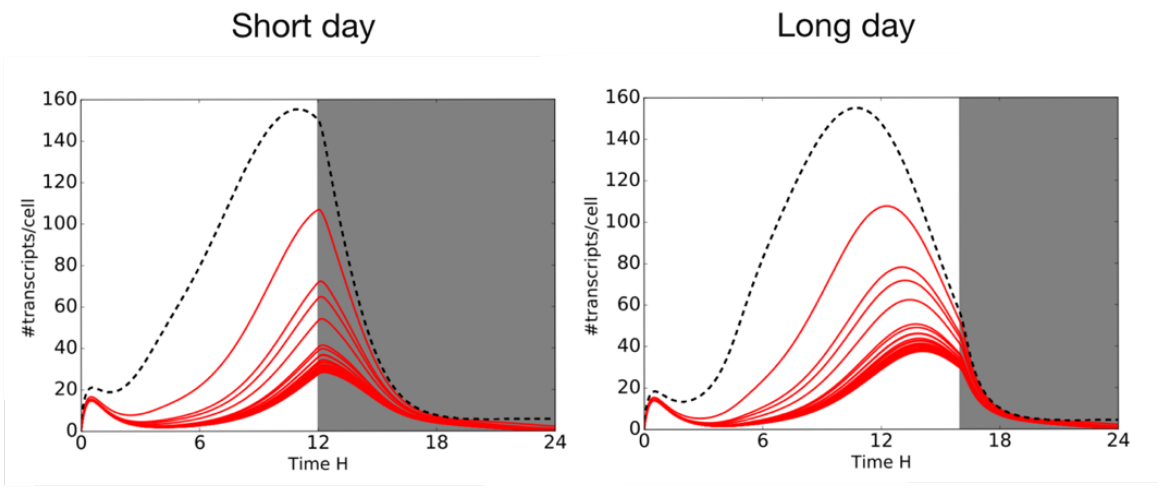


FIGURE 7.1 VARIATION IN THE DYNAMICS OF GI PREDICTED FROM INTEGRATING SEQUENCING DATA FORM THE 1001 GENOMES PROJECT AND A CLOCK MODEL IN ABSOLUTE UNITS U2017.2. ONLY 50 RANDOM ECOTYPES ARE PLOTTED FOR SIMPLICITY.

Finally, the clock model can start to incorporate local sequence information and use absolute units to assign affinities to the promoter regions of key target clock genes. The predicted affinities can be fed into the latest version of the clock model, together with local weather data, in order to understand the ecological and possibly evolutionary implications of variability in regulatory regions of clock promoters. The process illustrated above for *GI* might later be extended genome-wide, to consider clock-controlled target pathways it was done for PIF targets (Seaton et al. 2015) and hence multiple physiological traits. This direction shows the power of using Arabidopsis as a model system during the development of theoretical tools that might one day be used in ecology, or in analogous models for improving crops.

Interestingly, for the clock a lot of theoretical work has been produced because this system offers a bridge between disciplines and modellers.

References

Bar-Even, A. et al., 2011. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry*, 50(21), pp.4402–4410.

Buchler, N.E. & Cross, F.R., 2009. Protein sequestration generates a flexible ultrasensitive response in a genetic network. *Molecular systems biology*, 5(1), p.272.

Dai, S. et al., 2011. BROTHER OF LUX ARRHYTHMO is a component of the Arabidopsis circadian clock. *The Plant cell*, 23(3), pp.961–972.

de Montaigu, A. & Coupland, G., 2017. The timing of GIGANTEA expression during day/night cycles varies with the geographical origin of Arabidopsis accessions. *Plant signaling \& behavior*, 12(7), p.0.

Fogelmark, K. & Troein, C., 2014. Rethinking transcriptional activation in the Arabidopsis circadian clock. *PLoS computational biology*, 10(7), p.e1003705.

Fujiwara, S. et al., 2008. Post-translational regulation of the Arabidopsis circadian clock through selective proteolysis and phosphorylation of pseudo-response regulator proteins. *The Journal of biological chemistry*, 283(34), pp.23073–23083.

Hayama, R. et al., 2017. PSEUDO RESPONSE REGULATORs stabilize CONSTANS protein to promote flowering in response to day length. *The EMBO journal*, 36(7), pp.904–918.

Kemme, C.A., Esadze, A. & Iwahara, J., 2015. Influence of quasi-specific sites on kinetics of target DNA search by a sequence-specific DNA-binding protein. *Biochemistry*, 54(44), pp.6684–6691.

Murakami-Kojima, M. et al., 2002. The APRR3 component of the clock-associated APRR1/TOC1 quintet is phosphorylated by a novel protein kinase belonging to the WNK family, the gene for which is also

transcribed rhythmically in Arabidopsis thaliana. *Plant \& cell physiology*, 43(6), pp.675–683.

Nakamichi, N. et al., 2002. Compilation and characterization of a novel WNK family of protein kinases in Arabiodpsis thaliana with reference to circadian rhythms. *Bioscience, biotechnology, and biochemistry*, 66(11), pp.2429–2436.

Okada, M. et al., 2017. Synchrony of plant cellular circadian clocks with heterogeneous properties under light/dark cycles. *Scientific reports*, 7(1), p.317.

Piques, M. et al., 2009. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. *Molecular systems biology*, 5, p.314.

Seaton, D.D. et al., 2015. Linked circadian outputs control elongation growth and flowering in response to photoperiod and temperature. *Molecular systems biology*, 11(1), pp.776–776.

Sidaway-Lee, K. et al., 2014. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. *Genome biology*, 15(3), p.R45.

Somers, D.E., Kim, W.-Y. & Geng, R., 2004. The F-box protein ZEITLUPE confers dosage-dependent control on the circadian clock, photomorphogenesis, and flowering time. *The Plant cell*, 16(3), pp.769–782.

Wang, L., Fujiwara, S. & Somers, D., 2010. PRR5 regulates phosphorylation, nuclear import and subnuclear localization of TOC1 in the Arabidopsis circadian clock. *The EMBO journal*, 29(11), pp.1903–1915.

Wang, L., Kim, J. & Somers, D., 2013. Transcriptional corepressor TOPLESS complexes with pseudoresponse regulator proteins and histone deacetylases to regulate circadian transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 110(2), pp.761–766.

Wang, Y. et al., 2008. The plant WNK gene family and regulation of flowering time in Arabidopsis. *Plant biology (Stuttgart, Germany)*, 10(5), pp.548–562.

Wang, Z. & Tobin, E., 1998. Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. *Cell*, 93(7), pp.1207–1217.